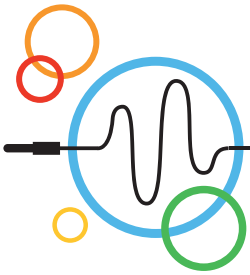
Modeling Expressed Emotions in Music using Pairwise Comparisons

Madsen, Jens; Nielsen, Jens Brehm; Jensen, Bjørn Sand; Larsen, Jan

Published in:
9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)

Publication date:
2012

Document Version
Publisher's PDF, also known as Version of record



Queen Mary
University of London

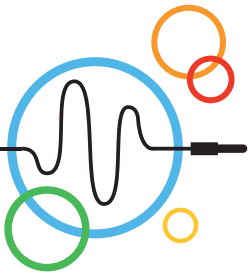
LMA

CMMR london 2012
Music & Emotions

Proceedings of the 9th International Symposium on Computer Music Modeling and Retrieval

19-22 June 2012

Queen Mary University of London



Cover graphic design by Céline Bokobza, www.boubokdesign.com
Proceedings design by Emmanouil Benetos, Dimitrios Giannoulis

ORGANISERS

PARTNERS

GOLD SPONSOR

SILVER SPONSOR

SPONSORS

Arts & Humanities
Research Council

Welcome to CMMR 2012

On behalf of the Conference Committee, it is a pleasure for us to welcome you to London for the 9th International Symposium on Computer Music Modeling and Retrieval (CMMR 2012): Music and Emotions. Jointly organised by the Centre for Digital Music, Queen Mary University of London, and the CNRS - Laboratoire de Mécanique et d'Acoustique, Marseille, France, CMMR 2012 brings together researchers, educators, librarians, composers, performers, software developers, members of industry, and others with an interest in computer music modeling, retrieval, analysis, and synthesis to join us for what promises to be a great event.

For this year's symposium, we chose the theme of Music and Emotions. Music can undoubtedly trigger various types of emotions within listeners. The power of music to affect our mood may explain why music is such a popular and universal art form. Research in cognitive science has investigated these effects, including the enhancement of intellectual faculties in given conditions by inducing positive affect. Music psychology has studied the production and discrimination of various types of expressive intentions and emotions in the communication chain between composer, performer and listener. Music informatics research has employed machine learning algorithms to discover relationships between objective features computed from audio recordings and subjective mood labels given by human listeners. But the understanding of the genesis of musical emotions and the mapping of musical variables to emotional responses remain complex research problems.

CMMR 2012 received over 150 submissions of papers, music, tutorials, and demos, and the committees chose the best of these to form a programme with seven technical sessions, two poster sessions, two panel sessions, a demo session, three concerts, two tutorials and a workshop. We are honoured to host the following invited speakers covering various aspects of our theme: Patrik Juslin (music psychology), Laurent Daudet (music signal processing) and Simon Boswell (film music composition). Ample time has been left between sessions for discussion and networking, complemented by the evening social programme, consisting of a welcome reception at Wilton's Music Hall, and a conference banquet on Thursday 21st June at Under the Bridge, which will feature a concert from the French band BBT and a jam session, in which delegates are invited to join in.

We wish to thank Mitsuko Aramaki, Richard Kronland-Martinet and Sølvi Ystad for giving us the opportunity to host this conference and for their work selecting the programme. We also thank our sponsors, who have generously supported the conference, allowing us to offset some of the costs of holding a conference in pre-Olympic London, including very busy scientific, musical and social programmes. Finally, we would like to take the opportunity to thank all of the members of the various committees, listed on the following pages, for their contribution to the symposium, the reviewers for their meticulous hard work, as well as the authors, presenters, composers and musicians taking part in the programme, without whom we would not have been able to host CMMR 2012.

We hope you enjoy the various scientific, musical and social events of the next four days, and that your time with us in London is rewarding.

Mathieu Barthet and Simon Dixon
CMMR 2012 Symposium Chairs

Organising Committee

Symposium Chairs

Mathieu Barthet, Centre for Digital Music, Queen Mary University of London
Simon Dixon, Centre for Digital Music, Queen Mary University of London

Proceedings Chairs

Richard Kronland-Martinet, CNRS-LMA (Marseille, France)
Solvi Ystad, CNRS-LMA (Marseille, France)
Mitsuko Aramaki, CNRS-LMA (Marseille, France)
Mathieu Barthet, Centre for Digital Music, Queen Mary University of London
Simon Dixon, Centre for Digital Music, Queen Mary University of London

Paper and Program Chairs

Richard Kronland-Martinet, CNRS-LMA (Marseille, France)
Mitsuko Aramaki, CNRS-LMA (Marseille, France)
Solvi Ystad, CNRS-LMA (Marseille, France)
Panos Kudumakis, Centre for Digital Music, Queen Mary University of London

Demonstrations, Panels & Tutorials Chairs

Daniele Barchiesi, Centre for Digital Music, Queen Mary University of London
Steven Hargreaves, Centre for Digital Music, Queen Mary University of London

Music Chairs and Concerts Curators

Andrew McPherson, Centre for Digital Music, Queen Mary University of London
Elaine Chew, Centre for Digital Music, Queen Mary University of London
Mathieu Barthet, Centre for Digital Music, Queen Mary University of London

Organising Committee

Daniele Barchiesi, Centre for Digital Music, Queen Mary University of London
Emmanouil Benetos, Centre for Digital Music, Queen Mary University of London
Luis Figueira, Centre for Digital Music, Queen Mary University of London
Dimitrios Giannoulis, Centre for Digital Music, Queen Mary University of London
Steven Hargreaves, Centre for Digital Music, Queen Mary University of London
Tom Heathcote, Queen Mary University of London
Sefki Kolozali, Centre for Digital Music, Queen Mary University of London
Sue White, Queen Mary University of London

Programme Committee

Mitsuko Aramaki, CNRS-LMA, France
Federico Avanzini, University of Padova, Italy
Isabel Barbancho, University of Málaga, Spain
Mathieu Barthet, Queen Mary University of London, UK
Roberto Bresin, KTH, Sweden
Marcelo Caetano, IRCAM, France
Antonio Camurri, University of Genova, Italy
Kevin Dahan, University of Paris-Est Marne-La-Vallée, France
Olivier Derrien, Toulon-Var University, France
Simon Dixon, Queen Mary University of London, UK
Barry Eaglestone, University of Sheffield, UK
George Fazekas, Queen Mary University of London, UK
Cédric Févotte, CNRS-TELECOM ParisTech, France
Bruno Giordano, McGill University, Canada
Emilia Gómez, Pompeu Fabra University, Spain
Goffredo Haus, Laboratory for Computer Applications in Music, Italy
Henkjan Honing, University of Amsterdam, The Netherlands
Kristoffer Jensen, Aalborg University, Denmark
Anssi Klapuri, Queen Mary University of London, UK
Richard Kronland-Martinet, CNRS-LMA, France
Panos Kudumakis, Queen Mary University of London, UK
Mark Levy, Last.fm, UK
Sylvain Marchand, Université de Bretagne Occidentale, France
Matthias Mauch, Queen Mary University of London, UK
Eduardo Miranda, University of Plymouth, UK
Marcus Pearce, Queen Mary University of London, UK
Emery Schubert, University of New South Wales, Australia
Björn Schuller, Munich University of Technology, Germany
Bob Sturm, Aalborg University, Denmark
George Tzanetakis, University of Victoria, Canada
Thierry Voinier, CNRS-LMA, France
Geraint A. Wiggins, Queen Mary University of London, UK
Sølvi Ystad, CNRS-LMA, France

Music Committee

Bertrand Arnold, Soundisplay, UK
Mathieu Barthet, Queen Mary University of London, UK
Elaine Chew, Queen Mary University of London, UK
Jacques Diennet, Ubris Studio, France
Philippe Festou, Laboratoire Musique et Informatique, France
Pascal Gobin, Conservatoire National de Région Marseille, France
Keeril Makan, Massachusetts Institute of Technology, USA
Ryan MacEvoy McCullough, Royal Conservatory, Canada
Andrew McPherson, Queen Mary University of London, UK
Eduardo Miranda, University of Plymouth, UK
Joo Won Park, Community College of Philadelphia, USA
Thomas Patteson, University of Pennsylvania, USA
Isaac Schankler, University of Southern California, USA
Jeff Snyder, Princeton University, USA
Dan Tidhar, King's College London, UK
Maurice Wright, Temple University, USA

Additional Reviewers

Samer Abdallah, Queen Mary University of London, UK
Emmanouil Benetos, Queen Mary University of London, UK
Charles Gondre, CNRS-LMA, France
Bas de Haas, Universiteit Utrecht, The Netherlands
Cyril Joder, Technische Universität München, Germany
Sefki Kolozali, Queen Mary University of London, UK
Andrew McPherson, Queen Mary University of London, UK
Martin Morrell, Queen Mary University of London, UK
Katy Noland, BBC, UK
Anaik Olivero, CNRS-LMA, France
Dan Tidhar, King's College, UK
Xue Wen, Queen Mary University of London, UK
Thomas Wilmering, Queen Mary University of London, UK
Massimiliano Zanoni, Politecnico di Milano, Italy

Programme

Tuesday 19th June

09:00-10:00	Registration
10:00-12:30	**Tutorial/Workshop 1** Pure Data and Sound Design (Andy Farnell)
10:00-12:30	**Tutorial 2** Musicology and Music Information Retrieval Tools (Daniel Leech-Wilkinson and Dan Tidhar)
10:00-12:30	**CMMR 2012 Music Concert** and C4DM Recording and Performance Spaces Tour
11:00-12:00	Coffee Break
12:30-13:30	Lunch
13:30-17:00	**Cross-Disciplinary Perspectives on Expressive Performance Workshop** Supported by the Arts and Humanities Research Council (AHRC)
15:00-17:00	Tour of British Library Sound Studios
15:00-16:00	Coffee Break
18:30-19:30	**Welcome Reception** (Balconies of Wilton's Grand Music Hall)
20:00-22:00	**New Resonances Festival** at Wilton's Music Hall (Concert 1)

Programme

Wednesday 20th June

09:00 - 09:30	Registration
09:30 - 09.45	Welcome and Announcements
09:45 - 10:45	**Keynote Talk 1**: "Hearing with our hearts: Psychological perspectives on music and emotions" (Prof. Patrik N. Juslin)
10:45 - 11:00	Coffee Break
11:00 - 12:40	**Oral session 1**: Music Emotion Analysis
12:40 - 13:00	**Yamaha Talk**
13:00 - 14:00	Lunch
14:00 - 15:00	**Poster Session 1**: Music Emotion: Analysis, Retrieval, and Multimodal Approaches, Synthesis, Symbolic Music-IR, Spatial Audio, Performance, Semantic Web
15:00 - 16:40	**Oral Session 2**: 3D Audio and Sound Synthesis
16:40 - 17:00	Coffee break
17:00 - 18:00	**Panel 1**: "Production Music: Mood and Metadata" **(**Dr. Mathieu Barthet, David Marston, Will Clark, Joanna Gregory, Marco Perry)
20:00 - 22:00	**New Resonances Festival** at Wilton's Music Hall (Concert 2)

Programme

Thursday 21st June

09:00 - 09:30	Registration
09:30 - 10:30	**Keynote Talk 2**: "The why, how, and what of sparse representations for audio and acoustics" (Prof. Laurent Daudet)
10:30 - 11:00	Coffee break
11:00 - 12:20	**Oral Session 3**: Computer Models of Music Perception and Cognition: Applications and Implications for MIR
12:20 - 12:40	**myfii Talk**
12:40 - 13:40	Lunch
13:40 - 15:00	**Poster session 2**: Computer Models of Music Perception and Cognition, Music Information Retrieval, Music Similarity and Recommendation, Musicology, Intelligent Music Tuition Systems
15:00 - 16:40	**Oral session 4**: Music Emotion Recognition
16:40 - 17:00	Coffee break
17:00 - 18:30	**Panel 2**: "The Future of Music Information Research" (Prof. Geraint A. Wiggins, Prof. Joydeep Bhattacharya, Prof. Tim Crawford, Dr. Alan Marsden, Prof. John Sloboda)
20:00 - 00:00	**Gala Dinner** at the Under The Bridge venue (Chelsea Football Club) followed by BBT Concert and Open Jam Session

Programme

Friday 22nd June

09:00 - 09:30	Registration
09:30 - 10:30	**Keynote Talk 3**: "Music In Cinema: How Soundtrack Composers Act On The Way People Feel" (Simon Boswell)
10:30 - 11:00	Coffee break
11:00 - 12:40	**Oral Session 5**: Music Information Retrieval
12:40 - 13:40	Lunch
13:40 - 14:40	**Demo Session**
13:40 - 14:40	**Yamaha Showcase**
14:40 - 15:40	**Session 6**: Film Soundtrack and Music Recommendation
15:40 - 16:00	Coffee break
16:00 - 17:20	**Oral Session 7**: Computational Musicology and Music Education
19:00 - 22:00	**New Resonances Festival** at Wilton's Music Hall (Concert 3)

Table of Contents

Preface i
Welcome iii
Organising Committee iv
Oral Session 1: Music Emotion Analysis 3
Continuous Response to Music using Discrete Emotion Faces
Emery Schubert, Sam Ferguson, Natasha Farrar, David Taylor, and Gary E. McPherson 3
Expressive dimensions in music
Tom Cochrane and Olivier Rosset 20
Emotion in Motion: A Study of Music and Affective Response
Javier Jaimovich, Niall Coghlan, and R. Benjamin Knapp 29
Psychophysiological measures of emotional response to Romantic orchestral music and their musical and acoustic correlates
Konstantinos Trochidis, David Sears, Dieu-Ly Tran, and Stephen McAdams 45
CCA and a Multi-way Extension for Investigating Common Components between Audio, Lyrics and Tags
Matt McVicar and Tijl De Bie 53
Poster Session 1: Music Emotion: Analysis, Retrieval, and Multimodal Approaches, Synthesis, Symbolic Music-IR, Spatial Audio, Performance, Semantic Web 70
Music Emotion Regression based on Multi-modal Features
Di Guan, Xiaoou Chen, and Deshun Yang 70
Application of Free Choice Profiling for the Evaluation of Emotions Elicited by Music
Judith Liebetrau, Sebastian Schneider, and Roman Jezierski 78
SUM: from Image-based Sonification to Computeraided Composition
Sara Adhitya and Mika Kuuskankare 94
Automatic Interpretation of Chinese Traditional Musical Notation Using Conditional Random Field
Rongfeng Li, Yelei Ding, Wenxin Li, and Minghui Bi 102
Music Dramaturgy and Human Reactions: Music as a Means for Communication
Javier Alejandro Garavaglia 112
ENP-Regex - a Regular Expression Matcher Prototype for the Expressive Notation Package
Mika Kuuskankare 128
The Role of Musical Features in the Perception of Initial Emotion
David Taylor, Emery Schubert, Sam Ferguson, and Gary McPherson 136
Sonic Choreography for Surround Sound Environments
Tommaso Perego 144
An Investigation of Music Genres and Their Perceived Expression Based on Melodic and Rhythmic Motifs
Debora C. Correa, F. J. Perez-Reche and Luciano da F. Costa 152

Subjective Emotional Responses to Musical Structure, Expression and Timbre Features: A Synthetic Approach
Sylvain Le Groux, Paul F. M. J. Verschure 160
Timing Synchronization in String Quartet Performance: A Preliminary Study
Marco Marchini, Panos Papiotis, and Esteban Maestre 177
Predicting Time-Varying Musical Emotion Distributions from Multi-Track Audio
Jeffrey Scott, Erik M. Schmidt, Matthew Prockup, Brandon Morton, and Youngmoo E. Kim 186
Codebook Design Using Simulated Annealing Algorithm for Vector Quantization of Line Spectrum Pairs
Fatiha Merazka 194
Pulsar Synthesis Revisited: Considerations for a MIDI Controlled Synthesiser
Thomas Wilmering, Thomas Rehaag, and André Dupke 206
Knowledge Management On The Semantic Web: A Comparison of Neuro-Fuzzy and Multi-Layer Perceptron Methods For Automatic Music Tagging
Sefki Kolozali, Mathieu Barthet, and Mark Sandler 220
Oral Session 2: 3D Audio and Sound Synthesis 233
A 2D Variable-Order, Variable-Decoder, Ambisonics based Music Composition and Production Tool for an Octagonal Speaker Layout
Martin J. Morrell and Joshua D. Reiss 233
Perceptual Characteristic and Compression Research in 3D Audio Technology
Ruimin Hu, Shi Dong, Heng Wang, Maosheng Zhang, Song Wang, and Dengshi Li 241
Rolling Sound Synthesis: Work In Progress
Simon Conan, Mitsuko Aramaki, Richard Kronland-Martinet, and Sølvi Ystad 257
EarGram: an Application for Interactive Exploration of Large Databases of Audio Snippets for Creative Purposes
Gilberto Bernardes, Carlos Guedes, and Bruce Pennycook 265
From Shape to Sound: Sonification of Two Dimensional Curves by Reenaction of Biological Movements
Etienne Thoret, Mitsuko Aramaki, Richard Kronland-Martinet, Jean-Luc Velay, and Sølvi Ystad 278
Oral Session 3: Computer Models of Music Perception and Cognition: Applications and Implications for MIR 287
The Role of Time in Music Emotion Recognition
Marcelo Caetano and Frans Wiering 287
The Intervalgram: An Audio Feature for Large-scale Melody Recognition
Thomas C. Walters, David A. Ross, and Richard F. Lyon 295
Perceptual Dimensions of Short Audio Clips and Corresponding Timbre Features
Jason Jiřì Musil, Budr Elnusairi, and Daniel Müllensiefen 311

Towards Computational Auditory Scene Analysis: Melody Extraction from Polyphonic Music
Karin Dressler 319
Poster Session 2: Computer Models of Music Perception and Cognition, Music Information Retrieval, Music Similarity and Recommendation, Musicology, Intelligent Music Tuition Systems 336
Predicting Emotion from Music Audio Features Using Neural Networks
Naresh N. Vempala and Frank A. Russo 336
Multiple Viewpoint Modeling of North Indian Classical Vocal Compositions
Ajay Srinivasamurthy and Parag Chordia 344
Comparing Feature-Based Models of Harmony
Martin Rohrmeier and Thore Graepel 357
Music Listening as Information Processing
Eliot Handelman and Andie Sigler 371
On Automatic Music Genre Recognition by Sparse Representation Classification using Auditory Temporal Modulations
Bob L. Sturm and Pardis Noorzad 379
A Survey of Music Recommendation Systems and Future Perspectives
Yading Song, Simon Dixon, and Marcus Pearce 395
A Spectral Clustering Method for Musical Motifs Classification
Alberto Pinto 411
Songs2See: Towards a New Generation of Music Performance Games
Estefanía Cano, Sascha Grollmisch, and Christian Dittmar 421
A Music Similarity Function Based on the Fisher Kernels
Jin S. Seo, Nocheol Park, and Seungjae Lee 429
Automatic Performance of Black and White n.2: The Influence of Emotions Over Aleatoric Music
Stefano Baldan, Adriano Baratè, and Luca A. Ludovico 437
The Visual SDIF interface in PWGL
Mika Kuuskankare 449
Application of Pulsed Melodic Affective Processing to Stock Market Algorithmic Trading and Analysis
Alexis Kirke and Eduardo Miranda 457
A Graph-Based Method for Playlist Generation
Debora C. Correa, Alexandre L. M. Levada, and Luciano da F. Costa . . 466
Compression-Based Clustering of Chromagram Data: New Method and Representations
Teppo E. Ahonen 474
GimmeDaBlues: An Intelligent Jazz/Blues Player And Comping Generator for iOS devices
Rui Dias, Telmo Marques, George Sioros, and Carlos Guedes 482
Oral Session 4: Music Emotion Recognition 492
Multidisciplinary Perspectives on Music Emotion Recognition: Implications for Content and Context-Based Models
Mathieu Barthet, György Fazekas, and Mark Sandler 492
A Feature Survey for Emotion Classification of Western Popular Music
Scott Beveridge and Don Knox 508

Support Vector Machine Active Learning for Music Mood Tagging
Álvaro Sarasúa, Cyril Laurier and Perfecto Herrera 518
Modeling Expressed Emotions in Music using Pairwise Comparisons
Jens Madsen, Jens Brehm Nielsen, Bjørn Sand Jensen, and Jan Larsen . 526
Relating Perceptual and Feature Space Invariances in Music Emotion Recognition
Erik M. Schmidt, Matthew Prockup, Jeffery Scott, Brian Dolhansky, Brandon G. Morton, and Youngmoo E. Kim 534
Oral Session 5: Music Information Retrieval 544
Automatic Identification of Samples in Hip Hop Music
Jan Van Balen, Martín Haro, and Joan Serrà 544
Novel Use of the Variogram for MFCCs Modeling
Simone Sammartino, Lorenzo J. Tardòn, and Isabel Barbancho 552
Automatic String Detection for Bass Guitar and Electric Guitar
Jakob Abeßer 567
Improving Beat Tracking in the Presence of Highly Predominant Vocals Using Source Separation Techniques: Preliminary Study
José R. Zapata and Emilia Gómez 583
Oracle Analysis of Sparse Automatic Music Transcription
Ken O'Hanlon, Hidehisa Nagano, and Mark D. Plumbley 591
Oral Session 6: Film Soundtrack and Music Recommendation 600
The Influence of Music on the Emotional Interpretation of Visual Contexts
Fernando Bravo 600
The Perception of Auditory-visual Looming in Film
Sonia Wilkie and Tony Stockman 611
Taking Advantage of Editorial Metadata to Recommend Music
Dmitry Bogdanov and Perfecto Herrera 618
Oral Session 7: Computational Musicology and Music Education 634
Bayesian MAP estimation of Piecewise Arcs in Tempo Time-series
Dan Stowell and Elaine Chew 634
Structural Similarity Based on Time-span Tree
Satoshi Tojo and Keiji Hirata 645
Subject and Counter-subject Detection for Analysis of the Well-Tempered Clavier Fugues
Mathieu Giraud, Richard Groult, and Florence Levé 661
Enabling Participants to Play Rhythmic Solos Within a Group via Auctions
Arjun Chandra, Kristian Nymoen, Arve Voldsund, Alexander Refsum Jenseniuz, Kyrre Glette, and Jim Torresen 674
Demo Session 691
Development of a Test to Objectively Assess Perceptual Musical Abilities
Lily Law and Marcel Zentner 691
Soi Moi...
n + n Corsino and Jacques Diennet 695
Author Index 697

Oral session 1:
Music Emotion Analysis

Continuous Response to Music using Discrete Emotion Faces

Emery Schubert[1], Sam Ferguson[2], Natasha Farrar[1], David Taylor[1] and Gary E. McPherson[3],

[1] Empirical Musicology Group, University of New South Wales, Sydney, Australia
[2] University of Technology, Sydney, Australia
[3] Melbourne Conservatorium of Music, University of Melbourne, Melbourne, Australia
E.Schubert@unsw.edu.au

Abstract. An interface based on expressions in simple graphics of faces were aligned in a clock-like distribution with the aim of allowing participants to quickly and easily rate emotions in music continuously. We developed the interface and tested it using six extracts of music, one targeting each of the six faces: 'Excited' (at 1 o'clock), 'Happy' (3), 'Calm' (5), 'Sad' (7), 'Scared' (9) and 'Angry' (11). 30 participants rated the emotion expressed by these excerpts on our 'emotion-face-clock'. By demonstrating how continuous category selections (votes) changed over time, we were able to show that (1) more than one emotion-face could be expressed by music at the same time and (2) the emotion face that best portrayed the emotion the music conveyed could change over time, and that the change could be attributed to changes in musical structure.

Keywords: Emotion in music, continuous response, discrete emotions, time-series analysis, film music.

1 Introduction

Research on continuous ratings of emotion expressed by music (that is, rating the music while it is being heard) has led to improvements in understanding and modeling music's emotional capacity. This research has produced time series models where musical features such as loudness, tempo, pitch profiles and so on are used as input signals which are then mapped onto emotional response data using least squares regression and various other strategies [1-4].

One of the criticisms of self-reported continuous response however, is the rating response format. During their inception in the 1980s and 1990s [5, 6] such measures have mostly consisted of participants rating one dimension of emotion (such as the happiness, or arousal, or the tension, and so on) in the music. This approach could be viewed as so reductive that a meaningful conceptualization of emotion is lost. For

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

example, Russell's [7, 8] work on the structure of emotion demonstrated that a large amount of variance in emotion can be explained by two fairly independent dimensions, frequently labeled valence and arousal. The solution to measuring emotion continuously can therefore be achieved by rating the stimulus twice (that is, in two passes), once along a valence scale (with poles of the scale labeled positive and negative), and once along an arousal scale (with poles labeled active and sleepy) [for another multi-pass approach see 9]. In fact, some researchers have combined these scales at right angles to form an 'emotion space' so as to allow a good compromise between reductive simplicity (the rating scale), and the richness of emotional meaning (applying what were thought to be the two most important dimensions in emotional structure simultaneously and at right angles) [e.g. 10, 11, 12].

The two dimensional emotion space has provided an effective approach to help untangle some of the relations between musical features and emotional response, as well as providing a deepening understanding of how emotions ebb and flow during the unfolding of a piece of music. However, the model has been placed under scrutiny on several occasions. The most critical matter that is of concern in the present research is theory and subsequent labeling of the emotion dimensions and ratings. For example, the work of Schimmack [13, 14] has reminded the research community that there are different ways of conceptualizing the key dimensions of emotion, and one dimension may have other dimensions hidden within it. Several researchers have proposed three key dimensions of emotion [15-17]. Also, dimensions used in the 'traditional' two dimensional emotion space may be hiding one or more dimensions. Schimmack demonstrated that the arousal dimension is more aptly a combination of underlying 'energetic arousal' and 'tense arousal'. Consider, for instance, the emotion of 'sadness'. On a single 'activity' rating scale with poles labeled active and sleepy, sadness will most likely occupy low activity (one would not imagine a sad person jumping up and down). However, in a study by Schubert [12] some participants consistently rated the word 'sad' in the high arousal region of the emotion space (all rated sad as being a negative valence word). The work of Schimmack and colleagues suggests that those participants were rating sadness along a 'tense arousal' dimension, because sadness does contain conflicting information about these two kinds of arousal – high tension arousal but low activity arousal.

Some solutions to the limitation of two dimensions are to have more than two passes when performing a continuous response (e.g. valence, tense arousal and activity arousal), or to apply a three dimensional GUI with appropriate hardware (such as a three dimensional mouse). However, in this paper we take the dilemma of dimensions as a point of departure and apply what we believe is the first attempt to use a discrete emotion response interface for continuous self-reported emotion ratings.

Discrete emotions are those that we think of in day-to-day usage of emotions, such as happy, sad, calm, energetic and so forth. They can each be mapped onto the emotional dimensions discussed above, but can also be presented as independent, meaningful conceptualizations of emotion [18-22]. An early continuous self-reported rating of emotion in music that demonstrated an awareness of this discrete structure was applied by Namba *et al*. [23], where a computer keyboard was labeled with fifteen different discrete emotions. As the music unfolded, participants pressed the key representing the emotion that the music was judged to be expressing at that time. The study has to our knowledge not been replicated, and we believe it is because the complexity of learning to decode a number of single letters and their intended emotion-word meaning. It seems likely that participants would have to shift focus between decoding the emotion represented on the keyboard, or finding the emotion and then finding its representative letter before pressing. And this needed to be done on the fly, meaning that by the time the response was ready to be made, the emotion in the music may have changed. The amount of training (about 30 minutes reported in the study) needed to overcome this cognitive load can be seen as an inhibiting factor.

Inspired by Namba *et al*'s pioneering work, we wanted to develop a way of measuring emotional response continuously but one which captured the benefits of discrete emotion rating, while applying a simple, intuitive user interface.

2 Using discrete facial expressions as a response format

By applying the work of some of the key research of emotion in music who have used discrete emotion response tools [24-26], and based on our own investigation [27], we devised a system of simple,

schematic facial expressions intended to represent a range of emotions that are known to be evoked by music. Further, we wanted to recover the topology of semantic relations, such that similar emotions were positioned beside one another, whereas distant emotions were physically more distant. This approach was identified in Hevner's [28-31] adjective checklist. Her system consisted of groups of adjectives, arranged in a circle in such a way as to place clusters of words near other clusters of similar meaning. For example, the cluster of words containing 'bright, cheerful, joyous …' was adjacent to the cluster of words containing 'graceful, humorous, light…', but distant from the cluster containing the words 'dark, depressing, doleful…'. Eventually, the clusters would form a circle, from which it derived its alternative names 'adjective clock' [32] and 'adjective circle' [31]. Modified version of this approach, using a smaller number of words, are still in use [33]. Our approach also used a circular form, but using faces instead of words. Consequently, we named the layout an 'emotion-face-clock'. Literate and non-literate cultures have become adept at speedy interpretation of emotional expression in faces [34, 35], making them more suitable for emotion rating tasks than words. Further, several emotional expressions are universal [36, 37] making the reliance on a non-verbal, non-language specific format appealing [38-40].

Selection of faces to be used for our response interface were based on the literature of commonly used emotion expressions to describe music [41], the recommendations made on a review of the literature by Schubert and McPherson [42] but also such that the circular arrangement was plausible. The faces selected corresponded roughly with the emotions from top moving clockwise (see Fig. 1): Excited (at 1 o'clock), Happy (3), Calm (5), Sad (7), Scared (9) and Angry (11 o'clock), with the bottom of the circle separated by Calm and Sad. The words used to describe the faces are selected for the convenience of the researchers. Although a circle arrangement was used, a small gap between the positive emotion faces and the negative emotion faces was imposed, because a spatial gap angry and excited, and between calm and sad reflected a semantic distance (Fig. 1). We did not impose our labels of the emotion-face expressions onto the participants. Pilot testing using retrospective ratings of music using the verbal expressions are reported in Schubert *et al*. [27].

3 Aim

The aim of the present research was to develop and test the emotion-face-clock as a means of continuously rating the emotion expressed by extracts of music.

4 Method

4.1 Participants

Thirty participants were recruited from a music psychology course that consisted of a range of students including some specializing in music. Self-reported years of music lessons ranged from 0 to 16 years, mean 6.6 years (SD = 5.3 years) with 10 participants reporting no music lessons ('0' years). Ages ranged from 19 to 26 years (mean 21.5 years, SD = 1.7 years). Twenty participants were male.

4.2 Software realisation

The emotion-face-clock interface was prepared, and controlled by MAX/MSP software, with musical extracts selected automatically and at random from a predetermined list of pieces. Mouse movements were converted into one of eight states: centre, one of the six emotions represented by schematic faces, and 'elsewhere' (Fig. 1). The eight locations were then stored in a buffer that was synchronized with the music, with a sampling rate of 44.1kHz. Given the redundancy of this sampling rate for emotional responses to music [which are in the order of 1 Hz – see 43], down-sampling to 25Hz was performed prior to analysis. The facial expressions moving around the clock in a clockwise direction were Excited, Happy, Calm, Sad, Scared and Angry. Note that the verbal labels for the faces are for the convenience of the researcher, and do not have to be the same as those used by participants. More important was that the expressions progressed sequentially around the clock such that related emotions were closer together than distant emotions, as described above. However, the quality of our labels were tested against participant data using the explicit labeling of the same stimuli in an earlier study [27].

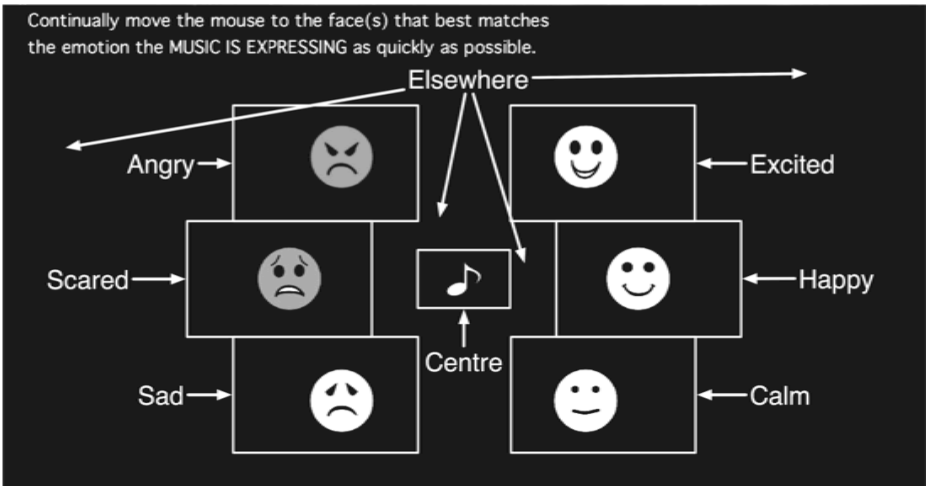


Fig. 1. Emotion-face-clock graphic user interface. This is a grayscale version. Face colours were yellow shades for right three faces (Excited [bright yellow], Happy and Calm), red for Angry, dark blue for Scared and light blue for Sad, based on [27]. Crotchet icon in Centre was green when ready to play, and grayed out, opaque when excerpt was playing. Text in top two lines provided instructions for the participant. White boxes, arrows and labels were not visible to the participants. These indicate the regions used to determine the eight response categories.

4.3 Procedure

Participants were tested one at a time. The participant sat at the computer display and wore headphones. After introductory tasks and instructions, the emotion-face-clock interface was presented, with a green icon (quaver) in the centre (Fig. 1). The participant was instructed to click the green button to commence listening, and to track the emotion that the music was expressing by selecting the facial expression that best matched the response. They were asked to make their selection as quickly as possible. When the participant moved the mouse over one of the faces, the icon of the face was highlighted to provide feedback. The participant was asked to perform several other tasks. The focus of the present report is on continuous rating over time of emotion that six extracts of music were expressing.

4.4 Stimuli

Because the aim of this study is to examine our new continuous response instrument, we selected six musical excerpts for which we had emotion ratings made using tradition post-performance ratings scales from a previous study [27]. The pieces were taken from Pixar animated movies, based on the principle that the music would be written to stereotypically evoke a range of emotions. The excerpts selected were 11 to 21 seconds long with the intention of primarily depicting each of the emotions of the six faces on the emotion-face-clock. In our reference to the stimuli in this report, they were labeled according to their target emotion: *Angry*, *Scared*, *Sad*, *Calm*, *Happy* and *Excited*. More information about the selected excerpts is shown in **Table 1**. When referring to a musical stimulus the emotion label is capitalized and italicised.

Table 1. Stimuli used in the study.

Stimulus code (target emotion)	Film music excerpt	Start time within CD track (MM'SS elapsed)	Duration of excerpt (s)
Angry	Up: 52 Chachki Pickup	00"53	17
Calm	Finding Nemo: Wow	00"22	16
Excited	Toy Story: Infinity and Beyond	00"15	16
Happy	Cars: McQueen and Sally	00"04	16
Sad	Toy Story 3: You Got Lucky	01"00	21
Scared	Cars: McQueen's Lost	00"55	11

5 Results and Discussion

Responses were categorized into one of eight possible responses (one of the six emotions, the centre location, and any other space on the emotion-face-clock labeled 'elsewhere' – see Fig. 1) based on mouse positions recorded during the response to each piece of music. This process was repeated for each sample (25 per second). Two main analyses were conducted. First, the relationships between the collapsed continuous ratings against rating scale results from a previous study using the same stimuli, and then an analysis of the time series responses for each of the six stimuli.

5.1 Summary responses

In a previous study, 26 participants provided ratings of each of the six stimuli used in the present study (for more details`, see [27] for details) along 11 point rating scales from '0 (not at all)' to '10 (a lot)'. The scales were labeled Angry, Scared, Sad, Calm, Happy and Excited. No faces were used in the response interface for that study.

The continuous responses from the current study were collapsed so that the number of votes a face received as the piece unfolded was tallied, producing a proportional representation of faces that were selected as indicating the emotion expressed by each face for a particular stimulus. The plots of these results are shown in Fig. 2. Take for example the responses made to the Angry excerpt. All participants first 'votes' were for the 'Centre' category because they had to click the icon at the centre of the emotion-face-clock to commence listening. As participants decided which face represented the emotion expressed, they moved the mouse to cover the appropriate face. So, as the piece unfolded, at any given time, some of the 30 participants might have the cursor on the Angry face, while some on the Scared face, and another who may not yet have decided remains in the centre or has moved the mouse, but not to a face ('elsewhere'). With a sampling rate of 25 Hz it was possible to see how these votes changes over time (the focus of the next analysis). At each sample, the votes were tallied into the eight categories. Hence each sample had a total of 30 votes (one per participant). At any sample it was possible to determine whether participants were or were not in agreement about the face that best represented the emotion expressed by the music.

The face by face tallies for each of these samples were accumulated and divided by the total number of samples for the excerpt. This provided a summary measure of the time-series to approximate the typical response profile for the stimulus in question. These profiles are reported in Fig. 2 in the right hand column. Returning to the *Angry* example we see that participants spent most time on the Angry face, followed by Scared and then the Centre. This suggests that the piece selected indeed best expressed anger according to the accumulated summary of the time series. The second highest votes belonging to the Scared face can be interpreted as a 'near miss' because of all the emotions on the clock, the scared face is semantically closest to the Angry face, despite obvious differences (for a discussion, see [27]). In fact, when comparing the accumulated summary with the post-

performance rating scale profile (from the earlier study), the time series produces a profile more in line with the proposed target emotion. The post-performance ratings demonstrate that Angry is only the third highest scored scale, after Scared and Excited. The important point, however, is that Scared and Excited are located on either side of the emotion-face-clock, making them the most semantically related alternatives to angry of the available faces. For each of the other stimuli, the contour of the profiles for post-performance ratings and accumulated summary of continuous response are identical.

These profiles matches are evidence for the validity of the emotion-face-clock because they mean that the faces are used to provide a similar meaning to the emotion words used in the post-performance verbal ratings. We can therefore be reasonably confident that at least five of the faces selected can be represented verbally by the five verbal labels we have used (the sixth – Anger, being confused occasionally with Scared). The similarity of the profile pairs in Fig. 2 is also indicative of the reliability of the emotion-face-clock because it more-or-less reproduces the emotion profile of the post-performance ratings.

Two further observations are made about the summary data. Participants spend very little time away from a face or the centre of the emotion-face-clock (the elsewhere region is selected infrequently for all six excerpts). While there is the obvious explanation that the six faces and the screen centre occupy the majority of the space on the response interface (see Fig. 1) the infrequent occurrence of the Elsewhere category also may indicate that participants are fairly certain about the emotion that the music is conveying. That is, when an emotion face is selected by a participant, they are likely to believe that to be the best selection, even if it is in disagreement with the majority of votes, or with the *a priori* proposed target emotion. If this were not the case, we might expect participants to hover in 'no mans land' of the emotion-face-clock—Elsewhere and Centre.

The 'no-mans-land' response may be reflected by the accumulated time spent on the centre category. As mentioned, time spent in the centre category is biased because participants always commence their responses from that region (in order to click the play button). The centre category votes can therefore be viewed as indicating two kinds of systematic responses: (1) initial response time and (2) response uncertainty. Initial response time is the time required for a participant to orient to the required task just as the temporally unfolding stimulus

commences. The orienting process generally takes several seconds to complete, prior to ratings becoming more 'reliable' [44-46]. So stimuli in Figure 2 with large bars for 'Centre' may require more time before an unambiguous response is made.

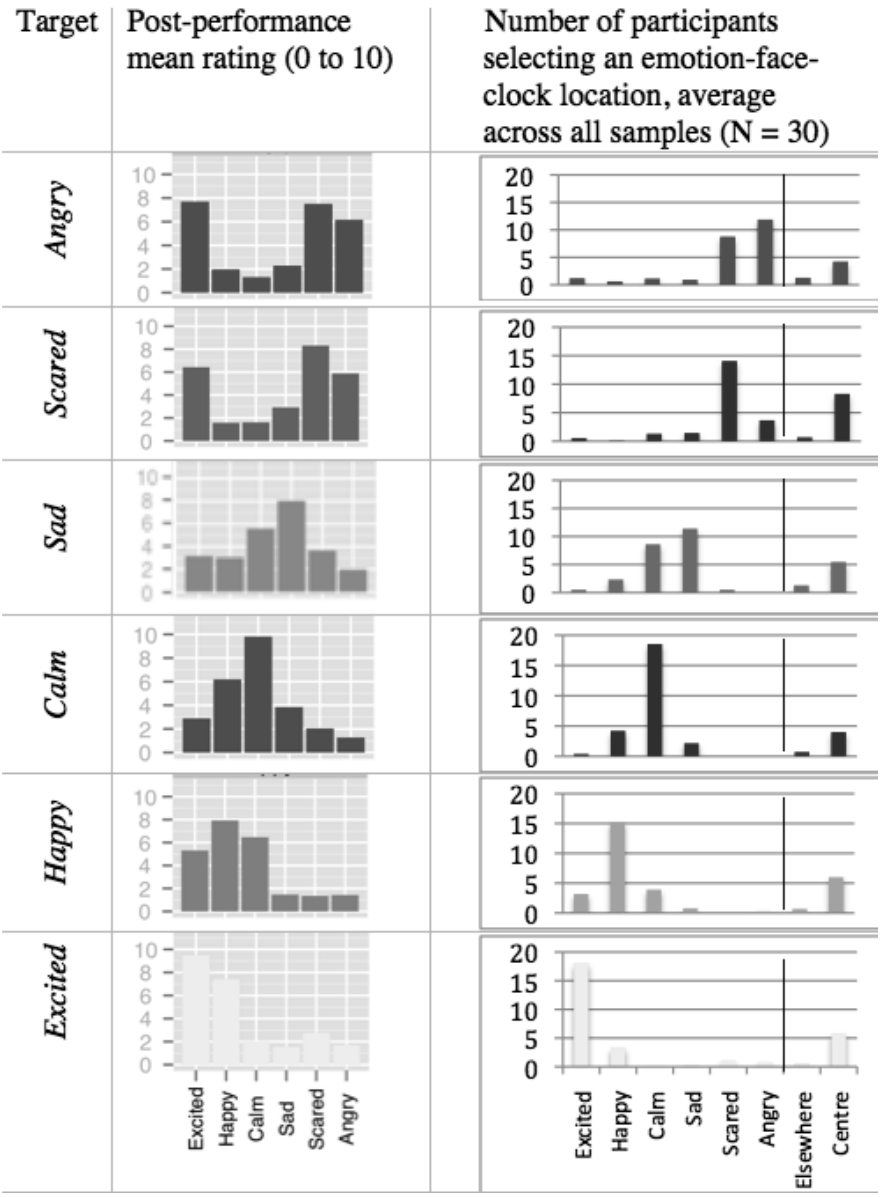


Fig. 2. Comparison of post performance ratings [from 27] (left column of charts) with sample averaged continuous response face counts for thirty participants (right column of charts) for the six stimuli, each with a target emotion shown in the leftmost column.

The relatively large amount of time spent in the Centre for this piece may, also, be an indicator of uncertainty of response. Well after a typical orientation period has passed, for this excerpt, uncertainty in rating remains (as will become clear in the next sub-section). The *Scared* stimulus has the largest number of votes for the Centre location (on average, at any single sample, eight out of thirty participants were in the centre of the emotion-face-clock). Without looking at the time series data, we may conclude that the *Scared* excerpt produced the least 'confident' rating, or that the faces provided were unable to produce satisfactory alternatives for the participants.

Using this logic (long time spent in the Centre and Elsewhere), we can conclude that the most confident responses were for those pieces where accumulated time spent in the Centre and Elsewhere were the lowest. The *Calm* stimulus had the highest 'confidence' rating (an average of about 4 participants at the Centre or Elsewhere combined). Interestingly, the *Calm* example also had the highest number of accumulated votes for any single category (the target, Calm emotion) — which was selected on average by 18 participants at any given time.

The analysis of summary data provides a useful, simple interpretation of the continuous responses. However, to appreciate the richness of the time-series responses, we now examine the time-series data for each stimulus.

5.2 Continuous responses

Fig. 3 shows the plots of the stacked responses from the 30 participants at each sample, for each stimulus. The beginning of each time series, thus, demonstrates that all participants commenced their response at the Centre (the first, left-most vertical 'line' of each plot is all black, indicating the Centre). By scanning for black regions for each of the plots in Fig. 2 some of the issues raised in the accumulated summary analysis, above, are addressed. We can see that the black and grey disappears for the *Calm* plot after 6 seconds have elapsed. For each of the other stimulus a small amount of doubt remains at certain times – in some cases a small amount of uncertainty is reported throughout (there are no samples in the *Scared* and *Excited* stimuli where all participants have selected a face). Further, the largest area of black and grey occurs in the *Scared* plot.

The time taken for 'most' participants to make a decision about the selection of a first face is fairly stable across stimuli. Inspection of Fig. 3 reveals that in the range of 0.5 seconds through to 5 seconds most participants have selected a phase. This provides a rough estimate of the initial orientation time for emotional response using categorical data (for more information`, see [44]).

Another important observation of the time-series of Fig. 3 is the ebb and flow of face frequencies. In the summary analysis it was possible to see when more than one emotion face was selected to identify the emotion expressed by the music. However, here we can see *when* these 'ambiguities' occur. The *Angry* and *Sad* stimuli provide the clearest examples of more than one dominant emotion. For the *Angry* excerpt, the 'Scared' face is frequently reported in addition to Angry. And the number of votes for the Scared face slightly increase toward the end of the excerpt. Thus, it appears that the music is expressing two emotions at the same time, or that the precise emotion was not available on the emotion-face-clock.

The *Sad* excerpt appears to be mixed with Calm for the same reasons (co-existence of emotions or precision of the measure). While the Calm face received fewer votes than the Sad face, the votes for Calm peak at around the 10th second (15 votes received over the time period 9.6 to 10.8s) of the *Sad* except. The excerpt is in a minor mode, opening with an oboe solo accompanied by sustained string chords and harp arpeggios. At around the 15th second (peaking at 18 votes over the time period 15.0 to 15.64s) the number of votes for Calm face begin to decrease and the votes for the Sad face peak. Hence, some participants find the orchestration and arch shaped melody in the oboe more calm than sad, until some additional information is conveyed in the musical signal (at around the 14th second), they remain on Calm. At the 10th second of this excerpt the oboe solo ends, and strings alone play, with cello and violin coming to the fore, with some *portamento* (sliding between pitches). These changes in instrumentation may have provided cues for participants to make the calm to sad shift after a delay of a few seconds [43].

Thus a plausible interpretation of the mixed responses is that participants have different interpretations of the various emotions expressed, *and* the emotion represented by the GUI faces. However, the changes in musical structure are sufficient to explain a change in response. What is important here, and as we have argued elsewhere, is that the difference between emotions is (semantically) small [27], and

that musical features could be modeled to predict the overall shift away from calmness and further toward sadness in this example.

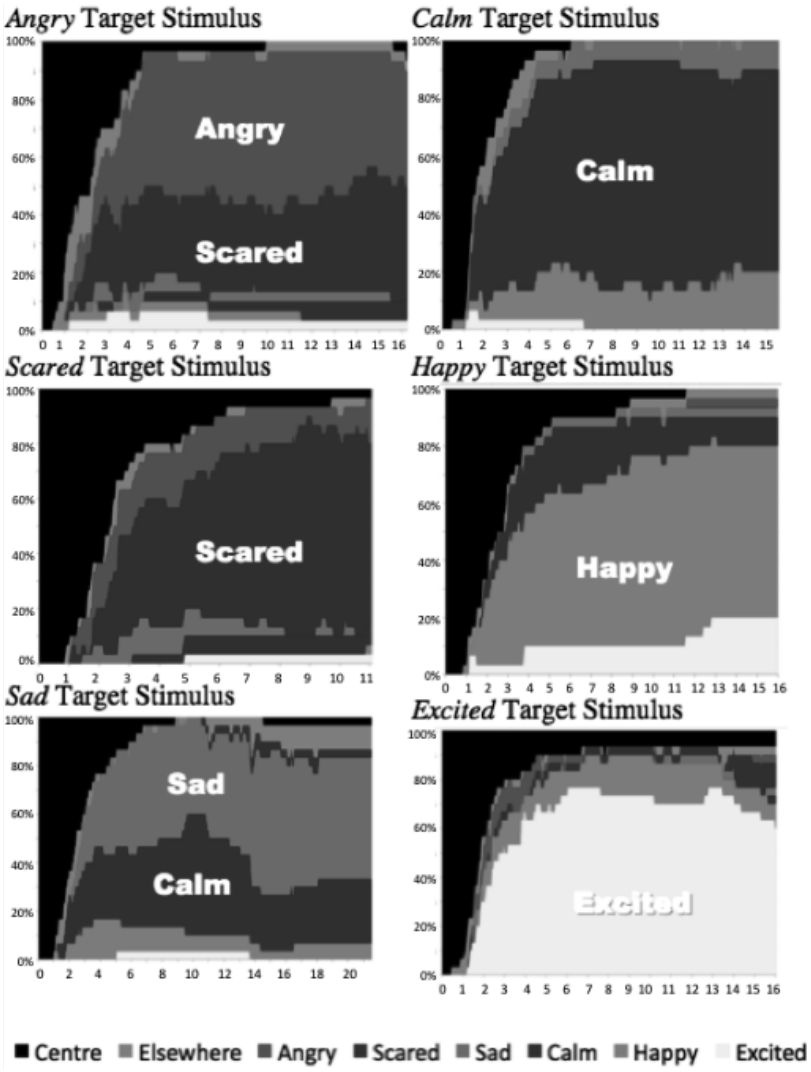


Fig. 3. Time series plots for each stimulus showing stacked frequency of faces selected over time (see **Table 1** for duration on x-axis) for the 30 participants (y-axis), with face selected represented by the colour code shown. Black and grey representing centre of emotion-face-clock (where all participants commence continuous rating task) and anywhere else respectively. Note that the most dominant colour (the most frequently selected face across participants and time) match with the target emotion of the stimulus.

6 Conclusions

In this paper we reported the development and testing of a categorical response interface consisting of a small number of salient emotional expressions upon which participants can rate emotions as a piece of music or other stimulus unfolds. We developed a small set of key emotional expression faces found in music research, and arranged them into a circle such that they were meaningfully positioned in space, and such that they resembled traditional valence-arousal rating scale interfaces (positive emotions toward the right, high arousal emotions toward the top). We called the response space an emotion-face-clock because the faces progressed around a clock in such a way that the expressions changed in a semantically related and plausible manner.

The interface was then tested using particular pieces that expressed the emotions intended to represent each of the six faces. The system was successful in measuring emotional ratings in the manner expected. The post-performance ratings used in an earlier study had profile contours that matched the profile contours of the accumulated summary of continuous response in the new device for all but the *Angry* stimulus. We took this as evidence for the reliability and validity of the emotion-face-clock as a self-report continuous measure of emotion.

Continuous response plots allowed investigation of the ebb and flow of ratings, demonstrating that for some pieces two emotions were dominant (the target *Angry* and target *Sad* excerpts in particular), but that the composition of the emotions changed over time, and that the change could be attributed to changes in musical features.

Further analysis will reveal whether musical features can be used to predict categorical emotions in the same way that valence/arousal models do (for a review, see [4]), or whether six emotion faces is optimal. Given the widespread use of categorical emotions in music metadata [47, 48], the categorical, discrete approach to measuring continuous emotional response is bound to be a fruitful tool for researchers interested in automating emotion in music directly into categorical representations.

Acknowledgments. This research was funded by the Australian Research Council (DP1094998).

References

1. Yang, Y.H., et al., *A regression approach to music emotion recognition*. Audio, Speech, and Language Processing, IEEE Transactions on, 2008. **16**(2): p. 448-457.
2. Schmidt, E.M., D. Turnbull, and Y.E. Kim. *Feature selection for content-based, time-varying musical emotion regression*. in *MIR '10 Proceedings of the international conference on Multimedia information retrieval*. 2010. ACM New York, NY.
3. Korhonen, M.D., D.A. Clausi, and M.E. Jernigan, *Modeling emotional content of music using system identification*. IEEE Transactions on Systems Man and Cybernetics Part B-Cybernetics, 2006. **36**(3): p. 588-599.
4. Schubert, E., *Continuous self-report methods*, in *Handbook of Music and Emotion: Theory, Research, Applications*., P.N. Juslin and J.A. Sloboda, Editors. 2010, OUP: Oxford. p. 223-253.
5. Madsen, C.K. and W.E. Frederickson, *The experience of musical tension: A replication of Nielsen's research using the continuous response digital interface*. Journal of Music Therapy, 1993. **30**(1): p. 46-63.
6. Nielsen, F.V., *Musical tension and related concepts*, in *The semiotic web '86. An international year-book*, T.A. Sebeok and J. Umiker-Sebeok, Editors. 1987, Mouton de Gruyter: Berlin: .
7. Russell, J.A., *Affective space is bipolar*. Journal of Personality and Social Psychology, 1979. **37**(3): p. 345-356.
8. Russell, J.A., *A circumplex model of affect*. Journal of Social Psychology, 1980. **39**: p. 1161-1178.
9. Krumhansl, C.L., *An exploratory study of musical emotions and psychophysiology*. Canadian Journal of Experimental Psychology, 1997. **51**(4): p. 336-352.
10. Cowie, R., et al., *FEELTRACE: An instrument for recording perceived emotion in real time*, in *Speech and Emotion: Proceedings of the ISCA workshop*, R. Cowie, E. Douglas-Cowie, and M. Schroeder, Editors. 2000, Co. Down.: Newcastle, UK. p. 19-24.
11. Nagel, F., et al., *EMuJoy: Software for continuous measurement of perceived emotions in music*. Behavior Research Methods, 2007. **39**(2): p. 283-290.
12. Schubert, E., *Measuring emotion continuously: Validity and reliability of the two-dimensional emotion-space*. Australian Journal of Psychology, 1999. **51**(3): p. 154-165.
13. Schimmack, U. and R. Rainer, *Experiencing activation: Energetic arousal and tense arousal are not mixtures of valence and activation*. Emotion, 2002. **2**(4): p. 412-417.
14. Schimmack, U. and A. Grob, *Dimensional models of core affect: A quantitative comparison by means of structural equation modeling*. European Journal Of Personality, 2000. **14**(4): p. 325-345.
15. Wundt, W., *Grundzüge der physiologischen Psychologie*. 1905, Leipzig: Engelmann.
16. Plutchik, R., *The emotions: Facts, theories and a new model*. 1962, New York: Random House. 204.
17. Russell, J.A. and A. Mehrabian, *Evidence for a 3-factor theory of emotions*. Journal of Research in Personality, 1977. **11**(3): p. 273-294.
18. Barrett, L.F. and T.D. Wager, *The Structure of Emotion: Evidence From Neuroimaging Studies*. Current Directions in Psychological Science, 2006. **15**(2): p. 79-83.
19. Barrett, L.F., *Discrete emotions or dimensions? The role of valence focus and arousal focus*. Cognition & Emotion, 1998. **12**(4): p. 579-599.
20. Lewis, M., J.M. Haviland-Jones, and L.F. Barrett, eds. Handbook of emotions (3nd ed.). 2008, The Guilford Press: New York, NY.
21. Izard, C.E., *The psychology of emotions*. 1991, NY: Plenum Press.
22. Izard, C.E., Organizational and motivational functions of discrete emotions, in Handbook of emotions, M. Lewis and J.M. Haviland, Editors. 1993, The Guilford Press: New York, NY. p. 631-641.

23. Namba, S., et al., *Assessment of musical performance by using the method of continuous judgment by selected description*. Music Perception, 1991. **8**(3): p. 251-275.
24. Juslin, P.N. and P. Laukka, *Communication of emotions in vocal expression and music performance: Different channels, same code?* Psychological Bulletin, 2003. **129**(5): p. 770-814.
25. Laukka, P., A. Gabrielsson, and P.N. Juslin, *Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion*. International Journal of Psychology, 2000. **35**(3-4): p. 288-288.
26. Juslin, P.N., *Communicating emotion in music performance: A review and a theoretical framework*., in *Music and emotion: Theory and research*, P.N. Juslin and J.A. Sloboda, Editors. 2001, Oxford University Press: London. p. 309-337.
27. Schubert, E., et al. *Sonification of Emotion I: Film Music*. in *The 17th International Conference on Auditory Display (ICAD-2011)*. 2011. Budapest, Hungary: International Community for Auditory Display (ICAD).
28. Hevner, K., *Expression in music: a discussion of experimental studies and theories*. Psychological Review, 1935. **42**: p. 187-204.
29. Hevner, K., *The affective character of the major and minor modes in music*. American Journal of Psychology, 1935. **47**: p. 103-118.
30. Hevner, K., *Experimental studies of the elements of expression in music*. American Journal of Psychology, 1936. **48**: p. 246-268.
31. Hevner, K., *The affective value of pitch and tempo in music*. American Journal of Psychology 49 1937, 621-630 Univ of Illinois Press, US, 1937.
32. Rigg, M.G., *The mood effects of music: A comparison of data from four investigators*. The journal of psychology, 1964. **58**(2): p. 427-438.
33. Han, B., et al. *SMERS: Music emotion recognition using support vector regression*. in *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2009)*. 2009. Kobe International Conference Center, Kobe, Japan: Kobe International Conference Center, Kobe, Japan, October 26-30, 2009.
34. Dimberg, U. and M. Thunberg, *Rapid facial reactions to emotional facial expressions*. Scandinavian Journal of Psychology, 1998. **39**(1): p. 39-45.
35. Britton, J.C., et al., *Facial expressions and complex IAPS pictures: common and differential networks*. Neuroimage, 2006. **31**(2): p. 906-919.
36. Waller, B.M., J.J. Cray Jr, and A.M. Burrows, *Selection for universal facial emotion*. Emotion, 2008. **8**(3): p. 435.
37. Ekman, P., *Facial expression and emotion*. American Psychologist, 1993. **48**(4): p. 384-392.
38. Lang, P.J., *Behavioral treatment and bio-behavioral assessment: Computer applications*, in *Technology in Mental Health Care Delivery Systems*, J.B. Sidowski, J.H. Johnson, and T.A. Williams, Editors. 1980, Ablex: Norwood, NJ. p. 119-137.
39. Bradley, M.M. and P.J. Lang, *Measuring Emotion - The Self-Assessment Mannequin And The Semantic Differential*. Journal Of Behavior Therapy And Experimental Psychiatry, 1994. **25**(1): p. 49-59.
40. Ekman, P. and E.L. Rosenberg, eds. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Series in affective science. 1997, Oxford University Press.: London.
41. Eerola, T. and J.K. Vuoskoski, *A comparison of the discrete and dimensional models of emotion in music*. Psychology of Music, 2011. **39**(1): p. 18-49.
42. Schubert, E. and G.E. McPherson, *The perception of emotion in music*, in *The child as musician: A handbook of musical development* G.E. McPherson, Editor. 2006, Oxford University Press: Oxford. p. 193-212.
43. Schubert, E., *Continuous measurement of self-report emotional response to music*, in *Music and emotion: Theory and research*, P.N. Juslin and J.A. Sloboda, Editors. 2001, Oxford University Press: Oxford. p. 393-414.

44. Schubert, E., *Reliability issues regarding the beginning, middle and end of continuous emotion ratings to music*. Psychology of Music, 2012.
45. Bachorik, J.P., et al., *Emotion in motion: Investigating the time-course of emotional judgments of musical stimuli*. Music Perception, 2009. **26**(4): p. 355-364.
46. Schubert, E. and W. Dunsmuir, *Regression modelling continuous data in music psychology*. , in *Music, Mind, and Science*, S.W. Yi, Editor. 1999, Seoul National University: Seoul. p. 298-352.
47. Trohidis, K., et al. *Multilabel classification of music into emotions*. in *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR 2008)*. 2008. Philadelphia, PA.
48. Levy, M. and M. Sandler. *A semantic space for music derived from social tags*. in *In Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007)*. 2007. Vienna, Austria.

Expressive dimensions in music

Tom Cochrane,[1] Olivier Rosset[2]

[1] Sonic Arts Research Centre, Queen's University Belfast
[2] Centre Interfacultaire en Sciences Affectives
thomas.cochrane@gmail.com

Abstract. This paper reports on an experiment into musical expressivity in which participants were asked to rate a number of short music pieces along three dimensions correlated with emotional states; valence, power and freedom. Results showed positive correlations between valence and the musical variables of dissonance and noise saturation, as well as between power and the musical variables of note sustain, tempo and reverb. More equivocal results were found for the dimension of freedom, and the musical variable of pitch height.

Keywords: Music, emotion, dimensions, valence, power, freedom, dissonance

1 Introduction

Continuous rating studies, in which participants are tasked to provide temporally continuous judgements of a phenomenon, are now frequently used to investigate correlations between musical features and emotional states [9], [11], [14], [18], (see also [6], [15] for reviews). Such studies have a distinct advantage over those which seek only summary judgements of expressive content since it is clear that the temporally varying nature of music is one of the main reasons it excels in the expression of emotions, and subtle variations as pieces progress can have dramatic effects on our sense of emotional content. It is a further advantage to utilize judgements of specific aspects of emotions, rather than simple emotion labels (i.e. 'happy', 'sad') both because such emotion labels are too general to guarantee their common understanding amongst multiple participants, and because pieces of music do not only become more or less sad, but reveal inflections on those emotions in much more subtle ways [17], [10]. As such the current study extends the continuous ratings approach to some new musical and emotional variables.

2 Emotional Dimensions

Typically continuous rating studies employ a model of emotions based on the two dimensions of valence and arousal, made popular by James Russell and Lisa Barrett [1], [13]. This model is widely recognized by psychologists and other scientists working in emotion related studies as a useful way to quantify emotions. Yet while we agree that the model is convenient to use, we note that it has several conceptual and practical limitations that encourage the use of alternative models, at least as a basis of comparison. Fontaine et al. [4] use factor analysis to show that at least 4 dimensions (which they label valence, arousal, power and certainty) must be used to

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

differentiate widely used emotion features (such as behavioural and feeling reports), while Cochrane [2] emphasizes that the dimensions of arousal and valence are not independent of one another, and therefore restrict the affective 'space' that can be effectively mapped. Moreover, valence and arousal are not capable of distinguishing even some of the most common negative emotions such as anger, panic and disgust.

Often a third dimension of 'power' or 'control' is recognized in emotion theories, and so this dimension has been used here as an alternative to arousal. Cochrane [2] also advocates the use of the dimension of 'freedom', since it is ambiguous whether power refers to a state of great strength or energy, or the ability to do what one wants. In common emotion episodes such as anger, these two aspects can be distinguished, since anger is typically a state of great strength or energy, while also involving a sense of constraint. In contrast, an emotion like joy may involve a sense of both great power and freedom. As such a dimension of freedom, specified here as the openness of the world to one's goals, is an important aspect of emotional experience and behaviour and may be usefully applied to our judgements of musical expressivity. At any rate, the use of additional dimensions against which to rate musical samples helps to provide a control for ratings along more popularly used dimensions.

Overall then, three dimensions of valence, power and freedom were employed in this study. Care was also taken to ensure that participants are provided with clear definitions of these dimensions (reproduced in table 1 below).

Table 1. Definitions of the emotion dimensions used in this study, as provided to participants.[1]

Dimension	Definition provided
Valence	By 'positive' and 'negative' character we mean whether the music sounds like a good/bad or pleasant/unpleasant feeling or seems to go with a situation that one would approach or avoid.
Freedom	By 'free' we mean whether the music sounds like a feeling of being able to do things or being open to the world as opposed to 'constrained' where one feels blocked or prevented from acting or shut off from the world.
Power	By 'powerful' and 'weak' we mean whether the music sounds like a feeling of energy/lack of energy or strength/weakness or seems to go with a situation where one is powerful or weak.

3 Method

A flash-based programme was developed by Olivier Rosset, enabling participants to provide continuous ratings of a short piece of music along a single specified dimension. Such a single dimension rating system contrasts with that of Nagel [12] and others, which tasks participants to rate two dimensions simultaneously, but which may not be practical when employing independent and non-standard dimensions, as in this study. Use of flash also allows participants to perform the study online. In this case participants used the computer mouse to adjust the vertical height of a line which

[1] These definitions were also translated into French by Kim Torres Eliard, though in the end only 3 francophone participants were enlisted. The translated text is available on request.

scrolled automatically across the page from left to right as the music played, and which was sampled every 250ms/4Hz. Seeing the line one produces affords the participant a clear sense of their overall judgement of the piece.

Meanwhile, a number of short pieces of music (typically around 30 seconds in length) were composed by Tom Cochrane using MIDI. These pieces were designed to systematically adjust one musical variable such as tempo or reverb while other variables were controlled. As much as possible, musical variables were adjusted in a linear fashion, resulting in a simple increase or decrease of the variable overall. Yet it should be noted that some variables, such as harmonic dissonance, can only be adjusted in a 'step-wise' fashion, while other variables such as reverb, though appearing to increase or decrease in a linear fashion within the confines of the MIDI sequencer, need not necessarily be perceived as such by participants.

For each musical variable, 3 pairs of pieces were provided; where as much as possible, a pair would preserve melodic and harmonic material while the desired musical variable was increased or decreased. Then between the 3 pairs of pieces, care was taken to vary the style, timbre and tonality of the music. For instance, one of the set may be based on a classical style melody, while another would employ a minimalist electronica style. Again (with the exception of pieces testing variances in harmonic dissonance) one piece may be largely in a minor key while another was in a major key, and others in more ambiguous tonalities. Such variation helps to justify the claim that the variables explored are capable of producing expressive variations that are independent of the specific musical context.

In all, 54 short music pieces were composed, designed to test 9 different musical variables. Each of these pieces were then rated on each of the three emotional dimensions of concern in this study. Since rating each sample on each dimension would be a prohibitively time-consuming and monotonous task, participants were randomly split into 3 groups. Each group would then rate 18 of the pieces on one dimension, followed by 18 on the second, then 18 on the third. The order in which pieces were presented within groups was randomized, as well as the order of dimensions ranked (though each dimension was rated as a block, in order for participants to be maximally aware of the nature of the dimension being judged). As mentioned above, very specific definitions for each dimension were provided, which were displayed throughout at the top of the rating window. Participants also had an opportunity to practice the rating task prior to each dimensional group.

Participants were recruited via online mailing lists such as <music-ir@listes.ircam.frm>, <auditory@lists.mcgill.ca> and online psychology forums such as ClinPsy.org.uk, in addition to psychology and philosophy students at Queen's University Belfast. This ensured a good mixture between musical experts and non-experts. The study was carried out between April and May 2011. Participants were not paid, and the results were anonymized. In total 50 participants were recruited, which when split into 3 groups ensured that each sample was rated on each dimension by at least 16 people.

3.1 Measuring Harmonic Dissonance

While most of the musical variables used in this study could be extracted in a fairly straightforward manner from the MIDI sequencer programme. Measuring the degree of harmonic dissonance in a piece requires a much more complex procedure of music information retrieval. In this case, MIDI note information was formatted using MATLAB to show all the notes simultaneously playing at any given moment. Each

interval between these notes was then attributed a dissonance score taken from the measure of sensory dissonance adapted for MATLAB by William Sethares [16]. This dissonance score was then multiplied by an additional dissonance score for each individual note playing at a given time. This additional individual-note measure is necessary because the sense of harmonic dissonance in an interval is relative to where that interval lies in relation to the tonic. For instance, in the key of C major, the major third between C and E natural is significantly less dissonant than the major third between Eb and G. This second dissonance score was adapted from the statistical frequency in which different tones appear in musical works, taken from Huron [7] based on the assumption justified by Huron that our sense of how well a given tone fits with the harmonic context determines the frequencies with which that tone tends to be employed in musical works.

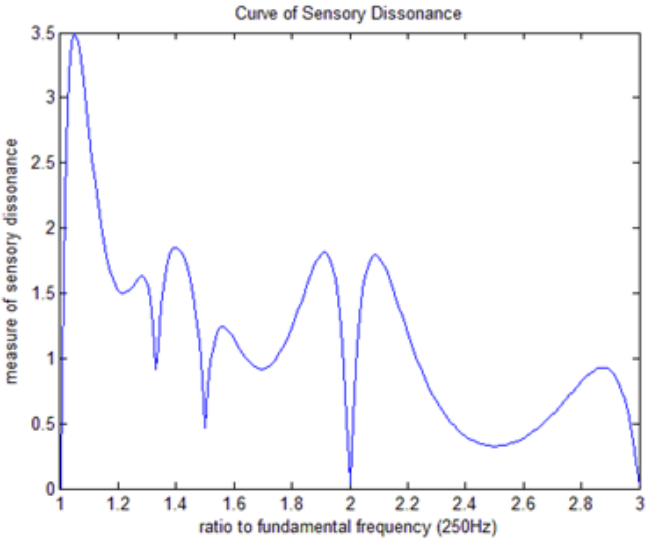


Fig. 3. Sensory Dissonance Curve from William Sethares [16]. The curve shows the level of perceived 'roughness' in tones relative to a fundamental frequency (in this case 250Hz). The relation of each tone to the fundamental frequency is expressed as a ratio. So in this case the number 2 on the x axis represents the octave above the fundamental, where 3 represents the fifth above that.

A further variable that was predicted in this study was that levels of harmonic dissonance would be correlated with the sense of emotional valence, but in a non-linear fashion. In particular, it was predicted that the technically most consonant intervals (e.g. simple octave) would be regarded as neutral in valence. The sense of valence should then trace an inverted U or Wundt curve, as dissonance approaches and then recedes from an optimally pleasant sense of harmony (say around a rich major chord). Theoretically, as dissonance moves towards an extreme where it becomes indistinguishable from noise, the sense of valence may again move back towards a neutral level. However, the kinds of tonal intervals used in this study would never approach that degree of complexity. As such, once a linear measure of dissonance was discerned for a piece of music, this was then transformed along the first five sixths of a sine curve- where the neutral starting point is assigned to sine0 and the maximal dissonance is assigned to sine1.572 i.e. 5 radians, the lowest point in

the sine curve. The difference in these two measures of dissonance are shown in figures 2 and 3 below, where it can be clearly seen that the transformed dissonance measure closely fits participants' judgements of valence in one of the pieces.

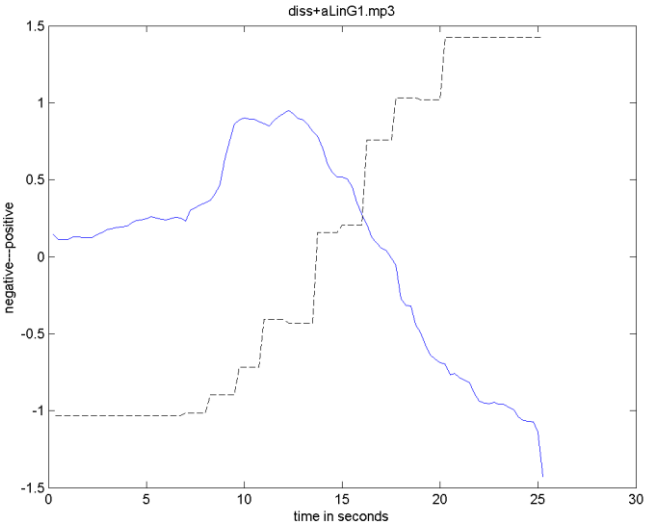


Fig. 2. Dissonance variable (*dotted line*) plotted linearly against participant ratings for valence (*solid line*). Both variables are normalized. Figure shows that as dissonance increases, ratings generally move towards the extreme negative, giving a fairly good negative correlation between the two variables.

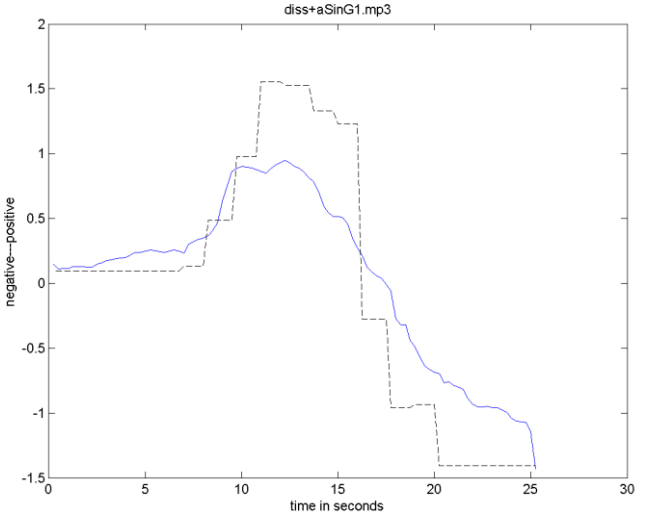


Fig. 3. Dissonance variable (*dotted line*) transformed along values of sine, where the neutral starting point is assigned to sine0 and the maximal dissonance is assigned to sine1.572 i.e. 5 radians, the lowest point in the sine curve.

4 Results

Table 2 below summarizes the results of the rating study for six musical variables of particular interest: harmonic dissonance, noise saturation, pitch height, note sustain, reverb and tempo. For each dimension, the six results are split into those in which the musical variable is increasing (e.g. greater dissonance) and those in which it is decreasing.

Table 2. Correlations (to two decimal places) between each musical variable, and each emotion dimension for 6 separate pieces of music (36 in total).

Music Variable	Dimension	Correlations					
		Variable Up			Variable Down		
Harmonic Dissonance	Valence[2]	0.95	0.77	0.82	-0.18	-0.37	-0.59
	Power	0.56	-0.76	0.21	-0.90	0.77	-0.57
	Freedom	-0.71	-0.74	-0.42	-0.47	0.58	0.80
Noise Saturation	Valence	-0.76	-0.89	-0.92	-0.47	-0.47	-0.22
	Power	0.97	-0.72	0.98	-0.80	0.64	0.30
	Freedom	0.10	-0.88	-0.13	-0.53	-0.84	0.02
Pitch Height	Valence	0.96	-0.31	0.92	-0.08	-0.55	-0.53
	Power	0.07	-0.70	0.76	-0.90	0.95	0.29
	Freedom	-0.64	-0.70	0.96	0.04	-0.91	-0.67
Note Sustain	Valence	-0.90	0.80	0.41	0.97	-0.99	-0.16
	Power	0.79	0.80	0.91	0.93	-0.43	0.91
	Freedom	-0.95	0.81	0.58	0.98	-0.82	0.73
Reverb	Valence	-0.91	-0.97	0.85	0.93	0.91	-0.97
	Power	-0.46	0.63	0.63	-0.89	-0.68	-0.96
	Freedom	-0.37	-0.94	0.42	0.92	0.90	-0.85
Tempo	Valence	-0.73	-0.84	0.92	0.91	0.34	-0.11
	Power	-0.49	0.68	0.68	0.99	0.90	0.94
	Freedom	-0.96	-0.84	0.98	0.89	0.92	-0.64

The correlations between dimensions and variables were then averaged for each set of 3 pieces to reveal which variables are most consistently correlated with which emotion dimensions.

[2] The dissonance measure here has been transformed along the values of sine, as detailed in section 3.1 above.

Table 3. Best averaged correlations (greater than +/- 0.8) across 3 pieces between musical variables and emotion dimensions. Correlations are shown to 3 decimal places.

Music Variable	Up/Down?	Dimension	Correlation	Significance
Harmonic Dissonance	Up	Valence	0.844	P<0.00001
Noise Saturation	Up	Valence	-0.858	P<0.00001
Reverb	Down	Power	-0.842	P<0.00001
Note Sustain	Up	Power	0.836	P<0.00001
Tempo	Down	Power[3]	0.946	P<0.00001

5 Discussion

The results show good correlations between dissonance, noise saturation and the dimension of valence, and between tempo, note sustain, reverb and the dimension of power. The connection between valence and dissonance replicates the findings of several other studies such as [3], [5] and [8] but also adds that the dissonance measure should be adjusted in a non-linear fashion. The expression of valence is further extended to the closely related variable of noise saturation. This is a plausible result since both musical features provide a sense of sensory roughness. Meanwhile, an increase in the sense of power is correlated with the *decrease* of reverb and the *increase* of note sustain. Reducing reverb results in a 'sharper' and 'closer' sound that may be psychologically associated with greater strength or energy,[4] while the increase of note sustain is a factor of loudness, that has already been associated with greater arousal [3]. Again, the correlation of a decrease in tempo with a decrease in power makes sense as a psychological connection between energy levels and one's speed of movement.

Ratings on the freedom dimension only produced good correlations on single pieces, and not as an average of three pieces in a set. In particular, while reverb was predicted to effect the sense of freedom, we see in table 2 that when reverb was decreasing, participants only agreed that it afforded a sense of greater constraint in two of the set of three, and they judged significantly in the *opposite* direction for the third piece. However, we also see that in some pieces of music, a significant rating for freedom is provided that is similar across both the increase and the decrease of the tested musical variable, indicating that some other musical variable in these pieces is consistently affecting the sense of freedom. These pieces are the second 'noise saturation' piece and to a lesser extent the second 'pitch height' piece. These pieces are both made from highly similar musical material which is slow and in a minor key- resulting in a fairly 'sad' sounding piece overall. This connection between the sense of freedom and sad sounding music would be worth investigating further. It was also noted that several pieces characterized by a high level of repetition resulted in low

[3] Note: as tempo goes down, power also goes down.

[4] Note that increased reverb can make a sound seem quieter, though as much as possible this possibility was controlled for in the composition of the reverb-varying pieces.

judgements of freedom, however these results were not very consistent. Again, it would be worth more systematically investigating a possible connection between the degree of repetition and the sense of freedom.

We also did not find strong correlations between pitch height and any emotion dimension. This musical variable was predicted to correlate with valence, but we see significant results only for two of the three pieces in the set. The most likely cause of this is that the second pitch height varying piece is in a minor key, and this sense of tonality probably overwhelmed any positive effect on valence that a rise in pitch may have achieved. The conflict between musical variables in this manner should also be more closely examined in future studies.

It is particularly notable that as a group, participants did not generally agree on the emotional significance of a musical variable where that variable was adjusted in both directions. For instance, where the increase in dissonance was widely agreed to result in a more 'negative' sound, the decrease of dissonance shows more confused results (in this case, one of the 3 pieces did not show a good correlation). Naturally, the fact that an average of the correlations across all judgements on a dimension, regardless of the direction of change, could not generally be obtained, must constrain any bald assertion that a certain musical feature goes with a certain emotion dimension. It is also possible that participants take time to adjust when a piece begins in a very dissonant mode, and only judge a mild sense of 'relief' when that dissonance is gradually removed. More generally however, it may be that adjustments in certain musical features are just more noticeable when they occur in one specific dimension. It would be worthwhile to test the above 'best' correlations in more detail, to explore more fluctuating levels in these musical variables within a longer piece of music.

Finally, it should be noted that while some good results were achieved in this study, there were a few limitations inherent to the experimental design. Naturally, a greater number of music pieces would enable a more robust measure of the dimension of concern. More importantly, listeners only had the chance to rate each piece once. While this would ensure a fresh response, it is not necessarily contrary to good judgements of expressive content to hear a piece over several times, and indeed to have the chance to return to one's judgement after hearing other examples.

Acknowledgments. This study was made possible by support for Tom Cochrane from the Swiss National Science Foundation, grant PBSKP1-130854 'The Mood Organ: Putting Theories of Musical Expression into Practice'.

References

1. Barrett, L. F. & Russell, J. A.: The Structure of Affect: Controversies and Emerging Consensus. Current Directions in Psychological Science, Vol. 8, No. 1, pp.10-14 (1999)

2. Cochrane, T.: Eight Dimensions for the Emotions. Social Science Information, Special issue, Vol. 48. No. 3 (2009)

3. Coutinho, E. & Cangelosi, A.: Musical Emotions: Predicting Second-by-Second Subjective Feelings of Emotion From Low-Level Psychoacoustic Features and Physiological Measurements. Emotion, (2011)

4. Fontaine, J., Scherer, K., Roesch, E. & Ellsworth, P.: The World of Emotions Is Not Two-Dimensional. Psychological Science, Vol. 18, No. 12, pp.1050-1057 (2007)

5. Fritz, T., Jentschke, S., Gosselin, N., Sammler, D., Peretz, I., Turner, R., Friederici, A. & Koelsch, S.: Universal Recognition of Three Basic Emotions in Music. Current Biology Vol. 19, No. 7, pp.573-576 (2009)

6. Gabrielsson, A. & Lindström, E.: The role of structure in the musical expression of emotions. Chapter 14 in Handbook of Music and Emotion: Theory, Research, Applications ed. P. Juslin and J. Sloboda. Oxford University Press, pp.467-400 (2010)

7. Huron, David: Sweet anticipation: Music and the psychology of expectation. Cambridge MA and London, MIT Press, (2006)

8. Juslin, P. & Laukka, P.: Communication of Emotions in Vocal Expression and Music Performance: Different Channels, Same Code? Psychological Bulletin Vol. 129, No. 5, pp.770–814 (2003)

9. Krumhansl, C. L.: An exploratory study of musical emotions and psychophysiology. *Canadian Journal of Experimental Psychology*, 51, pp.336–52 (1997)

10. Laukka, P., Juslin, P. & Bresin, R.: A dimensional approach to vocal expression of emotion. *Cognition & Emotion*, Vol. 19, No. 5, pp.633-653 (2005)

11. Madsen, C. K.: Emotion versus tension in Haydn's Symphony No. 104 as measured by the two-dimensional continuous response digital interface. *Journal of Research in Music Education*, 46, pp.546–54 (1998)

12. Nagel, F., Kopiez, R., Grewe, O. & Altenmüller, E.: 'EMuJoy' - Software for continuous measurement of perceived emotions in music: basic aspects of data recording and interface features. *Behavior Research Methods*, Vol. 39 No. 2, pp.283-290 (2007)

13. Russell, J. A.: A circumplex model of affect, *Journal of personality and social psychology* 39, pp.1161–78 (1980)

14. Schubert. E.: Modeling Perceived Emotion With Continuous Musical Features. *Music Perception*, Vol. 21, No. 4, pp.561–585 (2004)

15. Schubert, E.: Continuous Self-Report Methods. Chapter 9 in *Handbook of Music and Emotion, Theory, Research, Applications* ed. P. Juslin and J. Sloboda. Oxford University Press, pp.223-253 (2010)

16. Sethares, W.: Local consonance and the relationship between timbre and scale. *Journal of the Acoustic Society of America*, Vol. 94, No. 3 (September), pp. 1218-1228. Programme adapted from http://sethares.engr.wisc.edu/comprog.html (1993)

17. Vines, B. W, Krumhansl, C. L., Wanderley, M. M., Dalca, I. & Levitin, D. J.: Dimensions of Emotion in Expressive Musical Performance. *Ann. N.Y. Acad.* Sci. 1060, pp.1–5 (2005)

18. Vines, B. W., Nuzzo, R. L., & Levitin, D. J.: Analyzing temporal dynamics in music: Differential calculus, physics, and functional data analysis techniques. *Music Perception* 23, pp.137–52 (2005)

Emotion in Motion:
A Study of Music and Affective Response

Javier Jaimovich[1], Niall Coghlan[1] and R. Benjamin Knapp[2],

[1] Sonic Arts Research Centre, Queen's University Belfast
[2] Institute for Creativity, Arts, and Technology, Virginia Tech
{javier, niall, ben}@musicsensorsemotion.com

Abstract. 'Emotion in Motion' is an experiment designed to understand the emotional reaction of people to a variety of musical excerpts, via self-report questionnaires and the recording of electrodermal activity (EDA) and heart rate (HR) signals. The experiment ran for 3 months as part of a public exhibition, having nearly 4000 participants and over 12000 listening samples. This paper presents the methodology used by the authors to approach this research, as well as preliminary results derived from the self-report data and the physiology.

Keywords: Emotion, Music, Autonomic Nervous System, ANS, Physiological Database, Electrodermal Activity, EDR, EDA, POX, Heart Rate, HR, Self-Report Questionnaire.

1 Introduction

'Emotion in Motion' is an experiment designed to understand the emotional reactions of people during music listening, through self-report questionnaires and the recording of physiological data using on-body sensors. Visitors to the Science Gallery, Dublin, Ireland were asked to listen to different song excerpts while their heart rate (HR) and Electrodermal Activity (EDA) were recorded. The songs were chosen randomly from a pool of 53 songs, which were selected to elicit positive emotions (high valence), negative emotions (low valence), high arousal and low arousal. In addition to this, special effort was made in order to include songs from different genres, styles and eras. At the end of each excerpt, subjects were asked to respond to a simple questionnaire regarding their assessment of the song, as well as how it made them feel.

Initial analysis of the dataset has focused on validation of the different measurements, as well as exploring relationships between the physiology and the self-report data, which is presented in this paper.

Following on from this initial work we intend to look for correlations between variables and sonic characteristics of the musical excerpts as well as factors such as the effect of song order on participant responses and the usefulness of the Geneva Emotional Music Scale [1] in assessing emotional responses to music listening.

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

1.1 Music and Emotion

Specificity of musical emotions vs. 'basic' emotions. While the field of emotion research is far from new, from Tomkins theory of 'discrete' emotions [2] or Ekman's [3] studies on the 'universality' of human emotions to the fMRI enabled neuroimaging studies of today [4], there is still debate about the appropriateness of the existing 'standard' emotion models to adequately describe emotions evoked through musical or performance related experiences. It has been argued that many of the 'basic' emotions introduced by Ekman, such as anger or disgust, are rarely (if ever) evoked by music and that terms more evocative of the subtle and complex emotions engendered by music listening may be more appropriate [5]. It is also argued that the triggering of music-related emotions may be a result of complex interactions between music, cognition, semantics, memory and physiology as opposed to a direct result of audio stimulation [6, 7]. For instance a given piece of music may have a particular significance for a given listener e.g. it was their 'wedding song' or is otherwise associated with an emotionally charged memory.

While there is still widespread disagreement and confusion about the nature and causes of musically evoked emotions, recent studies involving real-time observation of brain activity seem to show that areas of the brain linked with emotion (as well as pleasure and reward) are activated by music listening [8]. Studies such as these would seem to indicate that there are undoubtedly changes in physiological state induced by music listening, with many of these correlated to changes in emotional state.

It is also important to differentiate between personal reflection of what emotions are expressed in the music, and those emotions actually felt by the listener [9]. In the study presented on this paper we specifically asked participants how the music made them feel as opposed to any cognitive judgments about the music.

During the last few decades of emotion research, several models attempting to explain the structure and causes of human emotion have been proposed. The 'discrete' model is founded on Ekman's research into 'basic' emotions, a set of discrete emotional states that he proposes are common to all humans; anger, fear, enjoyment, disgust, happiness, sadness, relief, etc. [10].

Russell developed this idea with his proposal of an emotional 'circumplex', a two or three axis space (valence, arousal and, optionally, power), into which emotional states may be placed depending on the relative strengths of each of the dimensions, i.e. states of positive valence and high arousal would lead to a categorization of 'joy'. This model allows for more subtle categorization of emotional states such as 'relaxation' [11].

The GEMS scale [1] has been developed by Marcel Zentner's team at the University of Zurich to address the perceived issue of emotions specifically invoked by music, as opposed to the basic emotion categories found in the majority of other emotion research. He argues that musical emotions are usually a combination of complex emotions rather than easily characterised basic emotions such as happiness or sadness. The full GEMS scale consists of 45 terms chosen for their consistency in describing emotional states evoked by music, with shorter 25 point and 9 point versions of the scale. These emotional states can be condensed into 9 categories which in turn group into 3 superfactors: vitality, sublimity and unease. Zentner also argues that musically evoked emotions are rare compared to basic/day-to-day emotions and that a random selection of musical excerpts is unlikely to trigger many

experiences of strong musically evoked emotions. He believes that musical emotions are evoked through a combination of factors which may include the state of the listener, the performance of the music, structures within the music, and the listening experience [5].

Lab vs. Real World. Many previous studies into musically evoked emotions have noted the difficulty in inducing emotions in a lab-type setting [12, 13], far removed from any normal music listening environment. This can pose particular problems in studies including measurements of physiology as the lab environment itself may skew physiological readings [14]. While the public experiment/installation format of our experiment may also not be a 'typical' listening environment, we believe that it is informal, open and of a non-mediated nature, which at the very least provides an interesting counterpoint to lab-based studies, and potentially a more natural set of responses to the stimuli.

1.2 Physiology of Emotion

According to Bradley and Lang, emotion has "almost as many definitions as there are investigators", yet "an aspect of emotion upon which most agree, however, is that in emotional situations, the body acts. The heart pounds, flutters stops and drops; palms sweat; muscles tense and relax; blood foils; faces blush, flush, frown, and smile" [15], page 581. A plausible explanation for this lack of agreement among researchers is suggested by Cacioppo et al. in [16], page 174. They claim that "...language sometimes fails to capture affective experiences - so metaphors become more likely vehicles for rendering these conscious states of mind", which is coherent with the etymological meaning of the word emotion; it comes from the Latin *movere*, which means to move, as by an external force.

For more than a century, scientists have been studying the relationship between emotion and its physiological manifestation. Analysis and experimentation has given birth to systems like the polygraph, yet it has not been until the past two decades, and partly due to improvements and reduced costs in physiological sensors, that we have seen an increase in emotion recognition research in scientific publications [17]. An important factor in this growth has been responsibility of the Affective Computing field [18], interested in introducing an emotion channel of communication to human computer interaction.

One of the main problems of emotion recognition experiments using physiology is the amount of influencing factors that act on the Autonomic Nervous System (ANS) [19]. Physical activity, attention and social interaction are some of the external factors that may influence physiological measures. This has led to a multi-modal theory for physiological differentiation of emotions, where the detection of an emotional state will not depend on a single variable change, but in recognizing patterns among several signals. Another issue is the high degree of variation between subjects and low repeatability rates, which means that the same stimulus will create different reactions in different people, and furthermore, this physiological response will change over time. This suggests that any patterns among these signals will only become noticeable when dealing with large sample sizes.

2 Methodology

2.1 Experimental Design

The aim of this study is to determine what (if any) are the relationships between the properties of an excerpt of music (dynamics, rhythm, emotional intent, etc.), the self-reported emotional response, and the ANS response, as measured through features extracted from EDA and HR. In order to build a large database of physiological and self-report data, an experiment was designed and implemented as a computer workstation installation to be presented in public venues. The experiment at the Science Gallery – Dublin[1] lasted for three months (June-August 2010), having nearly 4000 participants and over 12000 listening samples. The music selection included in its 53 excerpts contains a wide variety of genres, styles and structures, which, as previously mentioned, were selected to have a balanced emotional intent between high and low valence and arousal.

To be part of the experiment, a visitor to the Science Gallery was guided by a mediator to one of the four computer workstations, and then the individual followed the on-screen instructions to progress through the experiment sections (see Fig. 1 (b)). These would first give an introduction to the experiment and explain how to wear the EDA and HR sensors. Then, the participant would be asked demographic and background questions (e.g. age, gender, musical expertise, music preferences, etc.). After completing this section, the visitor would be presented with the first song excerpt, which was followed by a brief self-report questionnaire. The audio file is selected randomly from a pool of songs divided in the four affective categories. This was repeated two more times, taking each music piece from a different affective category, so each participant had a balanced selection of music. The visitor was then asked to choose the most engaging and the most liked song from the excerpts heard. Finally, the software presented the participant plots of his or her physiological signals against the audio waveform of the selected song excerpts. This was accompanied with a brief explanation of what these signals represent.

Software. A custom Max/MSP[2] patch was developed which stepped through the different stages of the experiment (e.g. instructions, questionnaires, song selection, etc.) without the need of supervision, although a mediator from the gallery was available in case participants had any questions or problems. The software recorded the participants' questions and physiological data into files on the computer, as well as some extra information about the session (e.g. date and time, selected songs, state of sensors, etc.). All these files were linked with a unique session ID number which was later used to build the database.

Sensors and Data Capture. MediAid POX-OEM M15HP[3] was used to measure HR using infra-red reflectometry, which detects heart pulse and blood oxygenation. The sensor was fitted by clipping on to the participant's fingertip as shown in Fig. 1 (a).

[1] http://www.sciencegallery.com/
[2] http://cycling74.com/products/max/
[3] http://www.mediaidinc.com/Products/M15HP_Engl.htm

(a) (b)

Fig. 1. (a) EDA and HR Sensors. (b) Participants during 'Emotion in Motion' experiment.

To record EDA, a sensor developed by BioControl Systems was utilized[4]. This provided a continuous measurement of changes in skin conductivity. Due to the large number of participants, we had to develop a 'modular' electrode system that allowed for easy replacement of failed electrodes.

In order to acquire the data from the sensors, an Arduino[5] microcontroller was used to sample the analogue data at 250 Hz and to send via serial over USB communication to the Max/MSP patch. The code from SARCduino[6] was used for this purpose. For safety purposes the entire system was powered via an isolation transformer to eliminate any direct connection to ground. Full frequency response closed-cup headphones with a high degree of acoustic isolation were used at each terminal, with the volume set at a fixed level.

Experiment Versions. During the data collection period, variations were made to the experiment in order to correct some technical problems, add or change the songs in the pool, and test different hypothesis. All of this is annotated in the database. For example, at the beginning participants were asked to listen to four songs in each session, later this was reduced to three in order to shorten the duration of the experiment. The questionnaire varied in order to test and collect data for different questions sets (detailed below), which were selected to compare this study to other experiments in the literature (e.g. the GEMS scales), analyse the effect of the questions in the physiology by running some cases without any questions, and also collect data for our own set of questions. The results presented in this paper are derived from a portion of the complete database with consistent experimental design.

Scales and Measures.

LEMtool The Layered Emotion Measurement Tool (LEMtool) is a visual measurement instrument designed for use in evaluating emotional responses to/with digital media [20]. The full set consists of eight cartoon caricatures of a figure expressing different emotional states (Joy/Sadness, Desire/Disgust,

[4] http://infusionsystems.com/catalog/product_info.php/products_id/203
[5] http://www.arduino.cc
[6] http://www.musicsensorsemotion.com/2010/03/08/sarcduino/

Fascination/Boredom, Satisfaction/Dissatisfaction) through facial expressions and body language. For the purposes of our experiment we used only the Fascination/Boredom images positioned at either end of a 5 point Likert item in which participants were asked to rate their levels of 'Engagement' with each musical excerpt.

SAM – Self Assessment Mannekin. The SAM is a non-verbal pictorial assessment technique, designed to measure the pleasure, arousal and dominance associated with a person's affective response to a wide range of stimuli [21]. Each point on the scale is represented by an image of a character with no gender or race characteristics, with 3 separate scales measuring the 3 major dimensions of affective state; Pleasure, Arousal, and Dominance. On the Pleasure scale the character ranges from smiling to frowning, on the Arousal scale the figure ranges from excited and wide eyed to a relaxed sleepy figure. The Dominance scale shows a figure changing in size to represent feelings of control over the emotions experienced.

After initial pilot tests we felt that it was too difficult to adequately explain the Dominance dimension to participants without a verbal explanation so we decided to use only the Pleasure and Arousal scales.

Likert Scales. Developed by the psychologist Rensis Likert [22], these are scales in which participants must give a score along a range (usually symmetrical with a mid-point) for a number of items making up a scale investigating a particular phenomenon. Essentially most of the questions we asked during the experiment were Likert items, in which participants were asked to rate the intensity of a particular emotion or experience from 1 (none) to 5 (very strong) or bipolar version i.e. 1 (positive) to 5 (negative).

GEMS – Geneva Emotional Music Scale. The 9 point GEMS scale [1] was used to ask participants to rate any instance of experiencing the following emotions: Wonder, Transcendence, Tenderness, Nostalgia, Peacefulness, Energy, Joyful activation, Tension, and Sadness. Again, they were asked to rate the intensity with which they were felt using a 5 point Likert scale.

Tension Scale. This scale was drafted by Dr. Roddy Cowie of QUB School of Psychology. It is a 5 point Likert scale with pictorial indicators at the Low and High ends of the scale depicting a SAM-type mannekin in a 'Very Relaxed' or 'Very Tense' state.

Chills Scale. This was adaptation from the SAM and featured a 5 point Likert scale with a pictorial representation of a character experiencing Chills / Shivers / Thrills / Goosebumps (CSTG), as appropriate, above the scale. The CSTG questions of the first version of the experiments were subsequently replaced with a single chills measure/question. For the purposes of the statistical analysis, the original results were merged using the mean to give a composite CSTG metric. This process was validated for consistency using both factor analysis and scale reliability test (Cronbach's Alpha of 7.83).

2.2 Song selection and description.

The musical excerpts used in the experiment were chosen by the researchers using several criteria: most were selected on the basis of having been used in previous experiments concerning music and emotion, while some were selected by the researchers for their perceived emotional content. All excerpts were vetted by the researchers for suitability. As far as possible we tried to select excerpts without lyrics, or sections in which the lyrical content was minimal[7].

Each musical example was edited down to approximately 90 seconds of audio. As much as possible, edits were made at 'musically sensible' points i.e. the end of a verse/chorus/bar. The excerpts then had their volume adjusted to ensure a consistent perceived level across all excerpts. Much of the previous research into music and emotion has used excerpts of music of around 30 seconds which may not be long enough to definitely attribute physiological changes to the music (as opposed to a prior stimulus). We chose 90 seconds duration to maximize the physiological changes that might be attributable to the musical excerpt heard. Each excerpt was also processed to add a short (< 0.5 seconds) fade In/Out to prevent clicks or pops, and 2 seconds of silence added to the start and end of each sound file. We also categorized each song according to the most dominant characteristic of its perceived affective content: Relaxed = Low Arousal, Tense = High Arousal, Sad=Low Valence, Happy = High Valence. Songs were randomly selected from each category pool every time the experiment was run with participants only hearing one song from any given category.

2.3 Feature extraction from physiology and database built

Database built. Once the signals and answer files were collected from the experiment terminals, the next step was to populate a database with the information of each session and listening case. This consisted in several steps, detailed below.

First, the metadata information was checked against the rest of the files with the same session ID number for consistency, dropping any files that had a wrong filename or that were corrupted. Subsequently, and because the clocks in each acquisition device and the number of samples in each recorded file can have small variations, the sample rates (SR) of each signal file were re-calculated. Moreover, some files had very different number of samples, which were detected and discarded by this process. To calculate the SR of each file, a MATLAB script counted the number of samples of each file, and obtained the SR using the duration of the song excerpt used in that recording. Two conditions were tested: a) that the SR was within an acceptable range of the original programmed SR (acquisition device) and b) that the SR did not present more than 0.5% variation over time. After this stage, the calculated SR was recorded as a separate variable in the database.

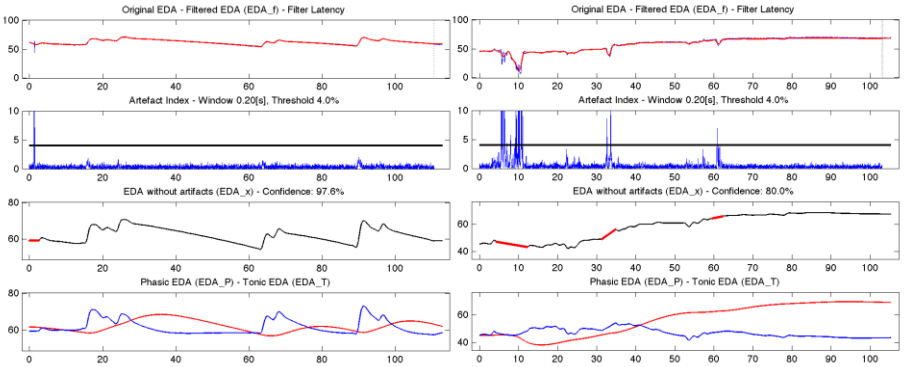
Finally, the data from each song excerpt was separated from its session and copied into a new case in the database. This means that each case in the database contains variables with background information of the participant, answers to the song questionnaire, and features extracted from the physiological signals, as well as

[7] The full list of songs used in the experiment may be found at http://www.musicsensorsemotion.com/demos/EmotionInMotion_Songs_Dublin.pdf

metadata about the session (experiment number, SR, order in which the song was heard, terminal number, date, etc.).

EDAtool and HRtool[8]**.** Two tools developed in MATLAB were used to extract features from the physiological data: EDAtool and HRtool. Extraction of features included detection and removal of artefacts and abnormalities in the data. The output from both tools consisted of the processed features vectors and an indication of the accuracy of the input signal, which is defined as the percentage of the signal which did not present artefacts. This value can be utilized later to remove signals from the database that fall below a specified confidence threshold.

EDAtool. EDAtool is a MATLAB function developed to pre-process the EDA signal. Its processing includes the removal of electrical noise and the detection and measurement of artefacts. Additionally, it separates the EDA signal into phasic and tonic components (please refer to [23] for a detailed description of EDA). Fig. 2 shows an example of the different stages of the EDAtool.



(a) (b)

Fig. 2. Stages of the EDAtool on a Skin Conductance signal. The top plots show the original signal (blue) and the low-passed filtered signal (red), which removes any electrical noise. The next plots show the artefact detection method, which identifies abrupt changes in the signal using the derivative. Fig. 2 (a) shows a signal above the confidence threshold used in this experiment, while signal in Fig. 2 (b) would be discarded. The 3rd row from the top shows the filtered signals with the artefacts removed. The bottom plots show the phasic (blue) and tonic (red) components of the signal.

HRtool. HRtool is a MATLAB function developed to convert the data from an Electrocardiogram (ECG) or Pulse Oximetry (POX) signal into an HR vector. This involves three main stages (see Fig. 3), which are the detection of peaks in the signal (which is different for a POX or an ECG signal), the measurement of the interval between pulses and the calculation of the corresponding HR value. Finally, the algorithm evaluates the HR vector replacing any values that are outside the ranges entered by the user (e.g. maximum and minimum HR values and maximum change ratio between two consequent pulses).

[8] http://www.musicsensorsemotion.com/tag/tools/

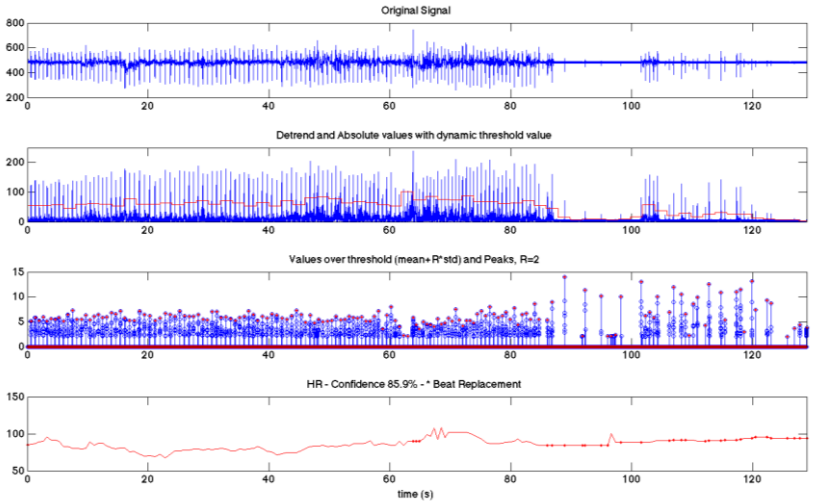


Fig. 3. Stages of the HRtool on an ECG signal. The top plot shows the raw ECG signal. The two middle plots show the peak detection stages, with a dynamic threshold. The bottom plot shows the final HR vector, with the resulting replacement of values that were outside the specified ranges (marked as dots in the plot). In this example, accuracy is at 85.9%, which falls below the acceptance tolerance for this experiment, and would be discarded as a valid case.

3. Preliminary Analysis

We are not aware of any similar study with a database of this magnitude, which has made it difficult to apply existing methodologies from smaller sized studies [17, 19]. Consequently, a large portion of the research presented in this paper has been dedicated to do exploratory analysis on the results; looking to identify relationships between variables and to evaluate the validity of the questionnaire and physiological measurements.

3.1 Preliminary results from questionnaire

General Demographic Information. After removing all data with artefacts, as described previously, an overall sample size of 3343 participants representing 11041 individual song listens was obtained. The remaining files were checked for consistency and accuracy and no other problems found.

The mean DOB was 1980 (Std. Dev. 13.147) with the oldest participants born in 1930 (22 participants, 0.2%). 47% of the participants were Male, 53% Female, with 62.2% identifying as 'Irish', and 37.8% coming from the 'Rest of the World'.

In the first version of the experiment participants heard four songs (1012 participants) with the subsequent versions consisting of three songs (2331 participants).

Participants were asked if they considered themselves to have a musical background or specialist musical knowledge, with 60.7% indicating 'No' and 39.3% indicating 'Yes'.

Interestingly, despite the majority of participants stating they had no specialist musical knowledge, when asked to rate their level of musical expertise from '1= No Musical Expertise' to '5= Professional Musician' 41.3% rated their level of musical expertise as '3'.

Participants were also asked to indicate the styles of music to which they regularly listen (by selecting one or more categories from the list below). From a sample of N=3343 cases, preferences broke down as follows: Rock 68.1%, Pop 60.3%, Classical 35%, Jazz 24.9%, Dance 34.2%, Hip Hop 27%, Traditional Irish 17%, World 27.9%, and None 1.2%.

Self-Report Data. An initial analysis was run to determine the song excerpts identified as most enjoyed and engaging. At the end of each experiment session, participants were asked which of the 3 or 4 (depending on experiment version) excerpts they had heard was the most enjoyable and which they had found most engaging. These questions appeared in all 5 versions of the experiment, making them the only ones to appear in all versions (other than the background or demographic questions).

The excerpts rated as 'Most Enjoyed' were James Brown 'Get Up (I Feel Like)' and Juan Luis Guerra 'A Pedir Su Mano' with these excerpts chosen by participants in 55% of the cases where they were one of the excerpts heard. At the other end of the scale, the excerpts rated lowest (fewest percentage of 'Most Enjoyed') were Slayer 'Raining Blood' and Dimitri Shostakovich 'Symphony 11, 0p. 103 - 2nd Movement' with these excerpts chosen by participants in 13% of the cases where they were one of the songs heard.

Participants were also asked to rate their 'Liking' of each excerpt (in experiment versions 1-3). Having analysed the mean values for 'Liking' on a per-song basis, the songs with the highest means were Jeff Buckley 'Hallelujah' (4.07/5) and The Verve 'Bittersweet Symphony' (4.03/5). The songs with the lowest mean values for 'Liking' were Slayer 'Raining Blood' (2.66/5) and The Venga Boys 'The Venga Bus is Coming' (2.93/5).

The excerpt rated most often as 'Most Engaging' was Clint Mansell's 'Requiem for a Dream Theme' with this excerpt chosen by participants in 53% of the cases where it was one of the excerpts heard. At the other end of the scale, the excerpt rated lowest (fewest percentage of 'Most Engaging') was Ceolteori Chualainn 'Marbhna Luimigh' with this excerpt chosen by participants in 11% of the cases where it was one of the excerpts heard.

Interestingly, when the mean values for 'Engagement' for each excerpt were calculated, Clint Mansell's 'Requiem for a Dream Theme' was only rated in 10th place (3.74/5), with Nirvana 'Smells Like Teen Spirit' rated highest (3.99/5), closely followed by The Verve 'Bittersweet Symphony' (3.95/5) and Jeff Buckley 'Hallelujah' (3.94/5). It was observed that while mean values for engagement are all within the 3-4 point range, there are much more significant differences between songs when participants were asked to rate the excerpt which they found 'Most Engaging', with participants clearly indicating a preference for one song over another.

The excerpts with the lowest mean values for 'Engagement' were Primal Scream 'Higher Than The Sun' (3.05/5) and Ceolteori Chualainn 'Marbhna Luimigh' (3.09/5). The excerpts with the highest mean values for Chills / Shivers / Thrills / Goosebumps (CSTG) were Jeff Buckley 'Hallelujah' (2.24/5), Mussorgsky 'A Night on Bare Mountain' (2.23/5) and G.A. Rossini 'William Tell Overture' (2.23/5). The excerpts with the lowest mean values for CSTG were Providence 'J.O. Forbes of Course' (1.4/5), Paul Brady 'Paddys Green Shamrock Shore' (1.43/5) and Neil Young 'Only Love Can Break Your Heart' (1.5/5).

An analysis was also run to attempt to determine the overall frequency of participants experiencing the sensation of CSTG. The number of instances where CSTG were reported as a 4 or 5 after a musical excerpt was tallied, giving 872 reports of a 4 or 5 from 9062 listens (experiment versions 1-3), meaning that significant CSTGs were experienced in around 10% of cases.

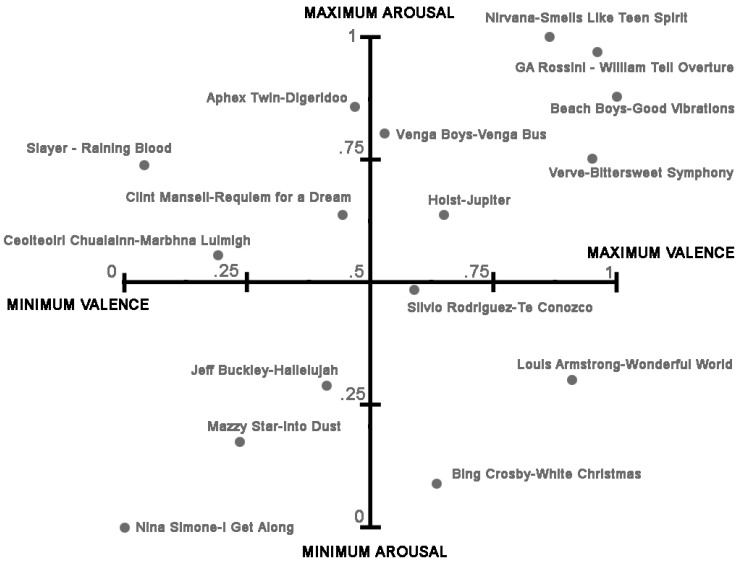


Fig. 4. Circumplex mapping of selected excerpts after a normalisation process to rescale the values 0 -1 with the lowest scoring excerpt in each axis as '0' and the highest as '1'.

A selection of the musical excerpts used (some of which were outliers in the above analyses) were mapped on to an emotional circumplex (as per Russell 1980), with Arousal and Valence (as measured using the SAM) as the Y and X axes respectively. An overall tendency of participants to report positive experiences during music listening was observed, even for songs which might be categorised as 'Sad' e.g. Nina Simone. Arousal responses were a little more evenly distributed but still with a slight positive skew. It seems that while some songs may be perceived as being of negative affect or 'sad', these songs do not in the majority of cases induce feelings of sadness. It may therefore be more appropriate to rescale songs to fit the circumplex from 'saddest' to 'happiest' (lowest Valence to highest Valence) and 'most relaxing' to 'most exciting' (lowest Arousal to highest Arousal) rather than using the absolute values reported (as seen on Fig. 4). This 'positive' skew indicating the rewarding

nature of music listening corroborates previous findings as documented in Juslin and Sloboda 2001 [24]. In future versions of this experiment we hope to identify songs that extend this mapping and are reported as even 'sadder' than Nina Simone.

3.2 Preliminary results from physiology

Features Extracted from Physiology. Due to the scope and nature of the experiment, the statistical analysis of the physiological signals has been approached as a continuous iteration, extracting a few basic features from the physiology, running statistical tests and using the results to extract new features. For this reason, the results from the physiology presented in this paper are still in a preliminary stage. The following features have been extracted from the 3 physiological vectors recorded in each case of the database (Phasic EDA, Tonic EDA and HR): Standard deviation of phasic EDA (*STD_EDAP*), mean of Phasic EDA (*mean_EDAP*), Tonic EDA final value divided by duration (*End_EDAT*), Tonic EDA trapezoidal numerical integration divided by duration (*Area_EDAT*), standard deviation of tonic EDA (STD_EDAT), difference between tonic EDA vector and linear regression of tonic start and end values (*Lin_EDAT*), mean of the 1st 10 raw EDA values (*Init_EDA*), mean HR (*HR*), mean heart rate variability (*mean_HRV*), HRV end value divided by duration (*End_HRV*), standard deviation of HRV (*STD_HRV*), square root of the mean squared difference of successive pulses (*RMSSD*), HRV low frequency (0.04-0.15Hz) component (*LF_HRV*), HRV high frequency (0.15-0.4Hz) component (*HF_HRV*) and ratio between *HF_HRV* and *LF_HRV* (*HtoL_HRV*).

Exploratory analysis and evaluation of measurements.
Dry skin issue. After removing any EDA signals that presented more than 10% of artefacts (measured by the EDAtool), preliminary analysis on several features extracted from the EDA vectors presented bimodality in the distribution of the features, which did not correspond to any of the variables measured or changed during the experiment (e.g. gender, age, song, etc.). The mean of the first 10 samples of each signal was calculated and added to the database in order to analyse the initial impedance of each subject. Fig. 5 shows the distribution of this variable.

The distribution shows a clear predominance of a group of participants that presented very high initial impedance (around the 160 mark). Although the origin of this irregularity is not clear, it is equivalent to the measurement of the EDA sensor when it has an open circuit (e.g. no skin connection). Due to the decision to use dry-skin electrodes (avoiding the application of conductive gel prior to the experiment), it is possible that this abnormality corresponds to a large group of participants in which the sensor did not make a good connection with the skin, probably due to them having a drier skin than the rest of the participants. It is also interesting to point out that there were a few hundred cases in which the sensor failed to work correctly (e.g. cases with conductivity near zero). For these reasons, the number of cases used for the analysis was filtered by the *Init_EDA* variable, looking for values that had normal impedance (above open-circuit value and below short-circuit value). This procedure solved the bimodality issue; at the cost of significantly reducing the valid cases in the database by 37%.

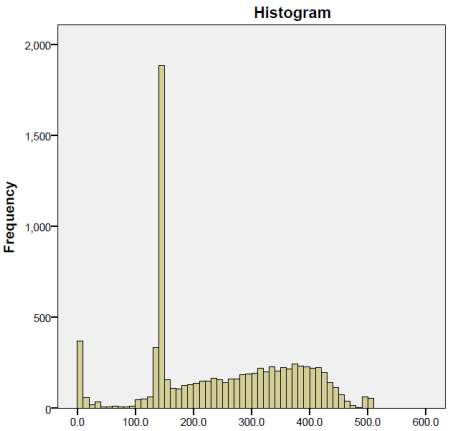


Fig. 5. Histogram of the mean of the 1st 10 samples of the EDA signal; equivalent to the initial conductivity. The histogram shows a large group of participants with an initial conductivity around the 160 mark (high impedance).

Correlation between physiological features and age. As expected, correlation between age and features extracted from physiology showed a negative relationship ($p < 0.01$ level, two-tailed) for several HR features (*STD_HRV*, *RMSSD*, *LF_HRV*, *HF_HRV*, *HtoL_HRV*), being the features that specify frequency components the ones with maximum correlation ($r < -0.4$).

Factor analysis of physiological features. Principal Component Analysis (PCA) was performed on a selection of features, excluding features with high degrees of correlation (it is important to state that all physiological features are derived from only two channels, EDA and HR, which can produce problems of multi-collinearity between features. This needs to be addressed prior to running a PCA). Principal Component Analysis shows three salient factors after rotation. These indicate a clear distinction between frequency-related features from HRV (Component 1: *STD_HRV*, *HF_HRV*, *LF_HRV*, *Age* and *RMSSD*), features from EDA (Component 2: *Area_GSRT*, *End_EDAT* and *STD_EDAP*) and secondary features from HRV (Component 3: *mean_HRV* and *End_HRV*).

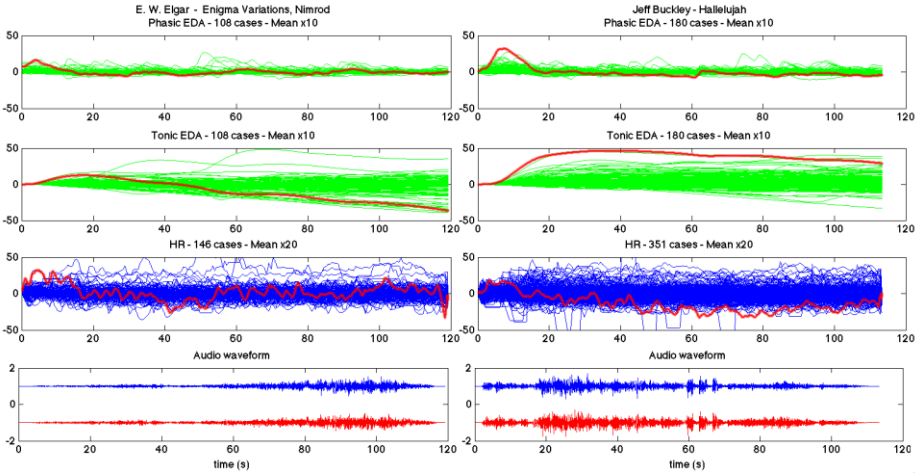
Correlation between factors and questionnaire. The three salient components from PCA were correlated against a selection of the self-report questionnaire: Song Engagement, Song Positivity, Song Activity, Song Tension, Song CSTG, Song Likeness and Song Familiarity. Results show a relationship between components 1 and 2 with the self-report questionnaire (see Table 1).

It is important to point out that the correlation coefficients presented below explain only a small portion of the variation in the questionnaire results. Furthermore, it is interesting that there was no significant correlation between CSTG and the 2nd component, taking into account that 10% of the participants reported to experience CSTG. Nevertheless, it is fascinating to see a relationship between physiological features and self-reports such as song likeness, positivity, activity and tension.

Table 1. Correlation between components from physiology and questionnaire

Question	Correlation by component (p<.001)		
	1	2	3
Song Engagement	-.081	.075	-
Song Positivity	-	.097	-
Song Activity	-	.110	-
Song Tension	-	.044	-
Song Chills/Shivers/Thrills/Goosebumps	-	-	-
Song Likeness	-.052	.061	-
Song Familiarity	-.060	.083	-

Music Dynamics vs. Physiology. Analysis of temporal changes in correlation with the excerpt's dynamic has been explored. Preliminary results show a relationship between the three physiological vectors; phasic EDA, tonic EDA and HRV, with changes in the music content, such as dynamics and structure. Fig. 6 shows two examples of pieces that present temporal correlation between physiology and music dynamic (a clear example is shown Fig. 6 (b) between the phasic EDA and the audio waveform after the 60 second mark).



(a) (b)

Fig. 6. Plots of changes in Phasic EDA, Tonic EDA, HR and audio waveform (top to bottom) during the duration of the song excerpt. Physiological plots show multiple individual responses overlapped, with the mean overlaid on top in red. Fig. 6 (a) plots are for Elgar's Enigma Variations, and plots in Fig. 6 (b) are for an excerpt of Jeff Buckley's Hallelujah.

4 Discussion

Due to the public gallery nature of this study, work has mainly been focused in improving the acquisition of signals, and the algorithms that correctly identify and remove noise and artefacts. Any unaccounted variation at this stage can impact the validity of the statistical tests that use physiological measurements. It is important to

point out that with the current sensor design, which requires no assistance and can be used by participants briefed with short instructions; we are obtaining approximately 65% valid signals (with a confidence threshold of 90%). This has to be taken into account when calculating group sizes for experiments that require physiological sensing of audiences.

The analysis of the physiological measures shows high levels of dispersion between participants for the same feature, which seems to indicate that large sample sizes need to be maintained for future experiments. Furthermore, a significant amount of the participants presented little to no variation in the features extracted from EDA. Nonetheless, the preliminary results presented in this paper are a significant indication of the possible relationships that explain the way we react to musical stimuli. Correlations between physiology and self-report questionnaire, in groups of this size, are a statement that this relationship undoubtedly exists. We are yet to further define the precise musical cues and variables that influence changes.

Next steps in the analysis will be focusing on additional physiological descriptors, multimodal analysis of the dataset, looking at temporal changes (versus the current whole song approach) and measures of correlation and entrainment with musical features. After the implementation in Dublin, 'Emotion in Motion' has been installed in public spaces in the cities of New York, Genoa and Bergen. Each iteration of the experiment has been enhanced and new songs have been added to the pool. We believe augmenting the sample size of these kinds of studies is a requirement to start elucidating the complex relationship between music and our affective response to it.

Acknowledgements. The authors would like to thank Dr. Miguel Ortiz-Perez for his invaluable contribution to the software design for this experiment, as well as Dr. Rodderick Cowie and Cian Doherty from QUB for their help with the questionnaire design. Finally, we would like to express our appreciation to the Science Gallery, Dublin for their support and funding of this experiment.

References

1. The Geneva Emotional Music Scales (GEMS) | zentnerlab.com, http://www.zentnerlab.com/psychological-tests/geneva-emotional-music-scales.
2. Tomkins, Tomkins, S.S.: Affect Imagery Consciousness - Volume II the Negative Affects. Springer Publishing Company (1963).
3. Ekman, P., Friesen, W.V.: The repertoire of nonverbal behavior: Categories, origins, usage, and coding. Semiotica. I, 49–98 (1969).
4. Salimpoor, V.N., Benovoy, M., Larcher, K., Dagher, A., Zatorre, R.J.: Anatomically distinct dopamine release during anticipation and experience of peak emotion to music. Nature Neuroscience. 14, 257–262 (2011).
5. Zentner, M., Grandjean, D., Scherer, K.R.: Emotions evoked by the sound of music: Characterization, classification, and measurement. Emotion. 8, 494–521 (2008).
6. Juslin, P.N., Västfjäll, D.: Emotional responses to music: the need to consider underlying mechanisms. Behav Brain Sci. 31, 559–575; discussion 575–621 (2008).

7. Balteş, F.R., Avram, J., Miclea, M., Miu, A.C.: Emotions induced by operatic music: Psychophysiological effects of music, plot, and acting: A scientist's tribute to Maria Callas. Brain and Cognition. 76, 146–157 (2011).
8. Trost, W., Ethofer, T., Zentner, M., Vuilleumier, P.: Mapping Aesthetic Musical Emotions in the Brain. Cerebral Cortex. (2011).
9. Gabrielsson, A., Juslin, P.N.: Emotional Expression in Music Performance: Between the Performer's Intention and the Listener's Experience. Psychology of Music. 24, 68–91 (1996).
10. Ekman, P.: An argument for basic emotions. Cognition & Emotion. 6, 169–200 (1992).
11. Russell, J.A.: A circumplex model of affect. Journal of Personality and Social Psychology. 39, 1161–1178 (1980).
12. Villon, O., Lisetti, C.: Toward Recognizing Individual's Subjective Emotion from Physiological Signals in Practical Application. Twentieth IEEE International Symposium on Computer-Based Medical Systems, 2007. CBMS '07. pp. 357–362. IEEE (2007).
13. Wilhelm, F.H., Grossman, P.: Emotions beyond the laboratory: Theoretical fundaments, study design, and analytic strategies for advanced ambulatory assessment. Biological Psychology. 84, 552–569 (2010).
14. Lantelme, P., Milon, H., Gharib, C., Gayet, C., Fortrat, J.-O.: White Coat Effect and Reactivity to Stress : Cardiovascular and Autonomic Nervous System Responses. Hypertension. 31, 1021 –1029 (1998).
15. Bradley, M.M., Lang, P.J.: Emotion and Motivation. Handbook of Psychophysiology. pp. 581–607 (2007).
16. Cacioppo, J.T., Bernston, G.G., Larsen, J.T., Poehlmann, K.M., Ito, T.A.: The Psychophysiology of Emotion. Handbook of Emotions. p. 173–91. Guilford Press (2000).
17. Kreibig, S.D., Wilhelm, F.H., Roth, W.T., Gross, J.J.: Cardiovascular, electrodermal, and respiratory response patterns to fear- and sadness-inducing films. Psychophysiology. 44, 787–806 (2007).
18. Picard, R.W.: Affective Computing. M.I.T Media Laboratory, Cambridge, MA (1997).
19. Kim, J., André, E.: Emotion Recognition Based on Physiological Changes in Music Listening. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 30, 2067–2083 (2008).
20. Huisman, G., Van Hout, M.: Using induction and multimodal assessment to understand the role of emotion in musical performance. Emotion in HCI – Designing for People. pp. 5–7. , Liverpool (2008).
21. Bradley, M.M., Lang, P.J.: Measuring emotion: The self-assessment manikin and the semantic differential. Journal of Behavior Therapy and Experimental Psychiatry. 25, 49–59 (1994).
22. Likert, R.: A technique for the measurement of attitudes. Archives of Psychology;Archives of Psychology. 22 140, 55 (1932).
23. Boucsein, W.: Electrodermal Activity. Springer (2011).
24. Juslin, P.N., Sloboda, J.A.: Music and Emotion: Theory and Research. Oxford University Press (2001).

Psychophysiological measures of emotional response to Romantic orchestral music and their musical and acoustic correlates

Konstantinos Trochidis, David Sears, Dieu-Ly Tran, Stephen McAdams

CIRMMT, Department of Music Research, McGill University
Konstantinos.Trochidis@mail.mcgill.ca, David.Sears@mail.mcgill.ca, Dieu-Ly.Tran@mail.mcgill.ca, smc@music.mcgill.ca

Abstract. This paper focuses on emotion recognition and perception in Romantic orchestral music. The study seeks to explore the relationship between perceived emotion and acoustic and physiological features. Seventy-five musical excerpts are used as stimuli to gather psychophysiological and behavioral responses of excitement and pleasantness from participants. A set of acoustic features ranging from low-level to high-level information was derived related to dynamics, harmony, timbre and rhythmic properties of the music. A set of physiological features based on blood volume pulse, skin conductance, facial EMGs and respiration rate measurements were also extracted. The feature extraction process is discussed with particular emphasis on the interaction between acoustical and physiological parameters. Statistical relations between audio, physiological features and emotional ratings from psychological experiments were systematically investigated. Finally, a step-wise multiple linear regression model is employed using the best features, and its prediction efficiency is evaluated and discussed. The results indicate that merging the acoustic and psychophysiological modalities substantially improves the emotion recognition accuracy.

Keywords: musical emotion, music perception, feature extraction, music information retrieval, psychophysiological response

1 Introduction

The nature of emotions induced by music has been a matter of much debate. Preliminary empirical investigations have demonstrated that basic emotions, such as happiness, anger, fear, and sadness, can be recognized in and induced by musical stimuli in adults and in young children [1]. The basic emotion model, which claims that music induces four or more basic emotions, is appealing to scientists for its empirical efficiency. However, it remains far from compelling for music theorists, composers, and music lovers because it is likely to underestimate the richness of emotional reactions to music that may be experienced in real life [2]. The question of whether emotional responses go beyond four main categories is a central issue for theories of human emotion [3]. An alternative approach to discrete emotions is to stipulate that musical emotions evolve continuously along two or three major psychological dimensions [4]. There are an increasing number of studies investigating

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

theoretical models in relation to music, the underlying factors and the mechanisms of emotional responses to music at behavioral [5, 6] and neurophysiological levels [7]. Many studies try to investigate the relationships between physiological features, such as electrocardiogram (ECG), electromyogram (EMG), skin conductance response (SCR) and respiration rate (RR), and emotional responses to music [9, 10, 11]. On the other hand, numerous studies explore the relationships between acoustic features and musical emotion [12, 13, 14]. Most of them try to extract a set of low- and high-level acoustical features representing various music descriptors (rhythm, harmony, tonality, timbre, dynamics) and correlate them with emotional ratings from participants.

The main aim of this paper is to implement an approach for music emotion recognition and retrieval based on both acoustic and physiological features. Our model is based on a previous study [15], which investigated the role of physiological response and peripheral feedback in determining the intensity and hedonic value of the emotion experienced while listening to music. Results from this study provide strong evidence that physiological arousal influences the intensity of emotion experienced with music and affects subjective feelings. Using this fusion model, we systematically combine structural features from the acoustic domain with psychophysiological features in order to further understand their relationship and the degree to which they affect subjective emotional qualities and feelings in humans.

2 Methods

2.1 Participants

Twenty non-musicians (M = 26 years of age) were recruited as participants (10 females). They reported less than 1 year of training on an instrument over the past five years, and less than two years of training in early childhood. In addition, all participants reported no hearing problems and that they liked listening to Classical and Romantic music.

2.2 Stimuli

Seventy-five musical excerpts from the late Romantic period were selected for the stimulus set. The selection criteria were as follows. The excerpts had to be anywhere from 35 to 45 seconds in duration, because we wanted 30 seconds of complete music after the fade-ins and fade-outs. The music was selected by the authors from the Romantic, late Romantic, or Neo-classical period (from 1815 to 1900). However, most excerpts were selected from the Romantic and late Romantic period. These genres were selected under the assumption that music from this period would elicit a variety of emotional reactions along both dimensions of the emotion model. Each excerpt had to clearly represent one of the four quadrants of the two-dimensional emotion space formed by the dimensions of arousal and valence. Ten excerpts were chosen from a previous study [16], 21 Romantic piano excerpts from [17] and 44 from our own personal selection. Aside from the high-arousal/negative-valence quadrant, which had 18 excerpts, the other three quadrants contained 19 excerpts. Moreover, the excerpts varied in orchestration, in order to explore the effect of timbre

variation on emotion judgments. Accordingly, there were 3 conditions: orchestral (24), chamber (26), and solo piano (25).

2.3 Procedure

We measured five different physiological signals for each of the participants: facial EMGs, skin conductance, respiration rate and blood volume pulse. The electrodes were placed on the following locations: the middle finger (BVP), the index and ring fingers (SC), above the zygomaticus muscle, located roughly in the center of the cheek (EMG), and above the corrugator super cilii muscle, located above the eyebrow (EMG). The respiration belt was placed around the torso in the middle of the rib cage just below the pectoral muscles.

Before beginning the experiment, a practice trial was presented to familiarize the participants with the experimental task. After listening to each musical excerpt, participants were asked to rate their level of experienced excitement and pleasantness on Likert scales.

3 Audio Feature Extraction

3.1 Low-Level acoustical features

A theoretical selection of musical features was made based on musical characteristics such as dynamics, timbre, harmony, register, and rhythm. A total of 100 features related to these characteristics were extracted from the musical excerpts. For all features, a series of statistical descriptors was computed such as the mean, the standard deviation and the linear slope of the trend across frames, i.e., the derivative. The MIR 1.3.4 Toolbox was used to compute the various low- and high-level descriptors [18].

3.1.1 Loudness features

We computed information related to the dynamics of the musical signals such as the RMS amplitude and the percentage of low-energy frames to see if the energy is evenly distributed throughout the signals or certain frames are more contrasted than others.

3.1.2 Timbre features

Mel Frequency Cepstral Coefficients (MFCCs) used for speech recognition and music modeling were employed. We derived the first 13 MFCCs. Another set of 4 features related to timbre were extracted from the Short-term Fourier Transform: spectral centroid, rolloff, flux, flatness entropy and spectral novelty which indicate whether the spectrum distribution is smooth or spiky. The size of the frames used to compute the timbre descriptors was 0.5 sec with an overlap of 50% between successive windows.

3.1.3 Tonality features

The signals were also analyzed according to their harmonic context. Descriptors such as the Chromagram (energy distribution of the signals wrapped in the 12 pitches), the key strength (i.e., the probability associated with each possible key candidate, through a cross-correlation with the Chromagram and all possible key candidates), the tonal Centroid (a vector derived from the Chromagram corresponding to the projection of the chords along circles of fifths or minor thirds) and the harmonic change detection function (flux of the tonal Centroid) were extracted.

3.1.4 Rhythmic features

A rhythmic analysis of the musical signals was performed. Descriptors such as the fluctuation (the rhythmic periodicity along auditory frequency channels) and the estimation of notes and number of onset and attack times per second were computed. Finally, the tempo of each excerpt in beats per minute (bpm) was estimated.

3.2 High-level acoustical features

In conjunction with the low-level acoustic descriptors, we used a set of high-level features computed with a slightly longer analysis window (3s). The high-level features are characteristics of music found frequently in music theory and music perception research.

3.2.1 Pulse Clarity

This descriptor measures the sensation of pulse in music. Pulse can be described as a fluctuation of musical periodicity that is perceptible as “beatings” in a sub-tonal frequency band below 20 Hz. The musical periodicity can be melodic, harmonic or rhythmic as long as it is perceived by the listener as a fluctuation in time [19].

3.2.2 Articulation

This feature attempts to estimate the articulation from musical audio signals by attributing to it an overall grade that ranges continuously from zero (staccato) to one (legato) by analyzing a set of attack times.

3.2.3 Mode

This feature refers to a computational model that rates excerpts on a bimodal major-minor scale. It calculates an overall output that varies along a continuum from zero (minor mode) to one (major mode) [14].

3.2.4 Event density

This descriptor measures the overall amount of simultaneous events in a musical excerpt. These events can be melodic, harmonic and rhythmic, as long as they can be

perceived as independent entities by listeners.

3.2.5 Brightness

This descriptor measures the sensation of how bright a musical excerpt is felt to be. Attack, articulation, or the unbalance or lacking of partials in other regions of the frequency spectrum can influence its perception.

3.2.6 Key Clarity

This descriptor measures the sensation of tonality, or tonal center in music. This is related to the sensation of how tonal an excerpt of music is perceived to be by listeners, disregarding its specific tonality, but focusing on how clear its perception is. This scale is also continuous, ranging from zero (atonal) to one (tonal).

4 Feature extraction of physiological signals

From the five psychophysiological signals we calculated a total of 60 features including conventional statistics in time series, frequency domain and sub-band spectra as suggested in [20].

4.1 Blood volume pulse

To obtain the HRV (heart rate variability) from the continuous BVP signal, each QRS complex was detected and the RR intervals (all intervals between adjacent R waves) or the normal-to-normal (NN) intervals (all intervals between adjacent QRS complexes resulting from sinus node depolarization) were determined. We used the QRS detection algorithm in [21] in order to obtain the HRV time series. In the time-domain of the HRV, we calculated statistical features including mean value, standard deviation of all NN intervals (SDNN), standard deviation of the first difference of the HRV, the number of pairs of successive NN intervals differing by greater than 50 ms (NN50), and the proportion derived by dividing NN50 by the total number of NN intervals. In the frequency-domain of the HRV time series, three frequency bands are of interest in general; very-low frequency (VLF) band (0.003-0.04 Hz), low frequency (LF) band (0.04-0.15 Hz), and high frequency (HF) band (0.15-0.4 Hz). From these sub-band spectra, we computed the dominant frequency and power of each band by integrating the power spectral densities (PSD) obtained by using Welch's algorithm, and the ratio of powers between the low-frequency and high-frequency bands (LF/HF).

4.2 Respiration

After detrending and low-pass filtering, we calculated the Breath Rate Variability (BRV) by detecting the peaks in the signal within each zero-crossing. From the BRV time series, we computed the mean value, SD, and SD of the first difference. In the

spectrum of the BRV, peak frequency, power of two sub-bands, low-frequency band (0-0.03Hz) and high-frequency band (0.03-0.15 Hz), and the ratio of power between the two bands (LF/HF) were calculated.

4.3 Skin conductance

The mean value, standard deviation, and mean of the first and second derivatives were extracted as features from the normalized SC signal and the low-passed SC signal using a 0.2 Hz cutoff frequency. To obtain a detrended SCR (skin conductance response) waveform without DC-level components, we removed continuous, piecewise linear trends in the two low-passed signals, i.e., very low-passed (VLP) with 0.08 Hz and low-passed (LP) signal with 0.2 Hz cutoff frequency.

4.4 Electromyography (EMGs)

For the EMG signals, we calculated similar types of features as in the case of the SC signal. From normalized and low-passed signals, the mean value of the entire signal, the mean of first and second derivatives, and the standard deviation were extracted as features. The number of occurrences of myo-responses and the ratio of these responses within VLP and LP signals were also added to the feature set in a similar manner used for detecting the SCR occurrence, but with 0.08 Hz (VLP) and 0.3 Hz (LP) cutoff frequencies.

5 Results

For the 75 excerpts a step-wise multiple linear regression to predict the participant ratings based on the acoustical and physiological descriptors between the acoustical and physiological descriptors and participant ratings were computed to gain insight into the importance of features for the arousal and valence dimensions of the emotion space. Table 1 provides the outcome of the MLR analysis of the acoustic features onto excitement and pleasantness coordinates of the excerpts and Table 2 the outcome of the analysis of the acoustic and physiological features onto the same coordinates. The resulting model provides a good account of excitement with an $R^2 = 0.81$ (see Table 1) using only the acoustic features spectral fluctuation ($\beta = 0.551$), entropy ($\beta = 0.302$) and spectral novelty ($\beta = –0.245$). For pleasantness, the model provides an $R^2 = 0.44$ using only the acoustic features Mode ($\beta = 0.5$), Key Clarity ($\beta = 0.27$) and entropy of Chroma ($\beta = 0.381$).

The model using both acoustic and physiological features provides an $R^2 = 0.85$ (see Table 2) with spectral fluctuation ($\beta = 0.483$), entropy ($\beta = 0.293$), spectral novelty ($\beta = –0.239$), the std of the first derivative of the zygomaticus EMG ($\beta = –0.116$), skin conductance ratio ($\beta = 0.156$), and the maximum value of the amplitude in blood volume pulse ($\beta = –0.107$). The model provides for pleasantness an $R^2 = 0.54$ using the acoustic and physiological features Mode ($\beta = 0.551$), Key Clarity ($\beta = 0.211$), entropy of Chroma ($\beta = 0.334$), the minimum of the std of the first derivative of the zygomaticus EMG ($\beta = 0.25$), and the minimum of the blood volume pulse ($\beta = -0.231$).

Table 1. Outcome of the multiple linear regression analysis of the acoustic features onto the coordinates of the emotion space.

Excitement	β	Pleasantness	β
Fluctuation	0.551	Mode	0.5
Enthropy	0.302	Key Clarity	0.27
Novelty	-0.245	Chroma Entropy	0.381

Table 2. Outcome of the multiple linear regression analysis using acoustic features and physiological features onto the coordinates of the emotion space.

Excitement	β	Pleasantness	β
Fluctuation	0.481	Mode	0.551
Enthropy	0.293	Key Clarity	0.221
Novelty	-0.23	Chroma Enthropy	0.334
1 diff EMGZ std	-0.11	1 diff EMGZ min	0.25
SC Ratio	-0.15	BVP min	-0.231

6 Conclusions

In the present paper, the relationships between acoustic and physiological features in emotion perception of Romantic music were investigated. A model based on a set of acoustic parameters and physiological features was systematically explored. The regression analysis shows that low- and high-level acoustic features such as Fluctuation, Entropy and Novelty combined with physiological features such as the first derivative of EMG Zygomaticus and Skin Conductance are efficient in modeling the emotional component of excitement. Further, acoustic features such as Mode, Key Clarity and the Chromagram combined with the minimum of the first derivative of EMG zygomaticus and blood volume pulse effectively model the emotional component of pleasantness. Using the existing approach merging acoustic and physiological features boosts the correlation with behavioral estimates of subjective feeling in listeners in terms of excitement and pleasantness. Results show an increase in the prediction rate of the model of 4% for excitement and 10% for pleasantness when psychophysiological measures are added to acoustic features.
Future work will explore and investigate by means of a similar model which low- and high-level acoustical and physiological features influence human judgments on semantic descriptions and perceptual qualities such as speed, articulation, harmony, timbre and pitch.

Acknowledgments. Konstantinos Trochidis was supported by a post-doctoral fellowship by the ACN Erasmus Mundus network. and a grant to Stephen McAdams from the Social Sciences and Humanities Research Council of Canada. The authors thank Bennett Smith for valuable technical assistance during the experiments.

References

1. Dolgin, K. G., & Adelson, E. H.: Age changes in the ability to interpret affect in sung and instrumentally-presented melodies. Psychology of Music, 18, 8--98 (1990)
2. Zentner, M., Grandjean, D., & Scherer, K.: Emotions evoked by the sound of music: Characterization, classification, and measurement. Emotion, 8, 494--521 (2008)
3. Ekman, P.: The nature of emotion: Fundamental questions. New York: Oxford University Press (1994)
5. Russell, J. A.: A circumplex model of affect. Journal of Personality and Social Psychology, 39, 1161--1178 (1980)
6. Juslin, P. N., & Västfjäll, D.: Emotional responses to music: The need to consider underlying mechanisms. Behavioral and Brain Sciences, 31, 559--575 (2008)
7. Juslin, P. N., & Sloboda, J. A.: Psychological perspectives on music and emotion. In: P. N. Juslin & J. A. Sloboda (eds.), Music and emotion: Theory and research (pp. 361--392). New York: Oxford University Press (2001)
8. Schmidt L A., Trainor, L. J.: Frontal brain activity (EEG) distinguishes valence and intensity of musical emotions, Cognition and Emotion, 15, 487—500 (2001)
9. Gomez, P., & Danuser, B.: Relationships between musical structure and psychophysiological measures of emotion. Emotion, 7(2), 377--387 (2007)
10. Khalfa, S., Peretz, I., Blondin, J.P., & Manon, R.: Event-related skin conductance responses to musical emotions in humans. Neuroscience Letters, 328, 145--149 (2002)
11. Sears, D., Ogg, M., Benovoy, M., Tran, D. L., S. McAdams, S.: Predicting the Psychophysiological Responses of Listeners with Musical Features. Poster presented at the 51st Annual Meeting of the Society for Psychophysiological Research, Boston, MA, September 14-18 (2011)
12. Eurola, T. ,Lartillot, O. ,Toiviainen, P.: Prediction of Multidimensional Emotional ratings in Music from Audio Using Multivariate Regression Models, in Proc. ISMIR (2009)
13. Fornari, J. & Eerola, T.: Computer Music Modeling and Retrieval. Genesis of Meaning in Sound and Music, in Lecture Notes in Computer Science, chapter The Pursuit of Happiness in Music: Retrieving Valence with Contextual Music Descriptors, 5493, 119-133. Springer (2009)
14. Saari, P., Eerola, T., & Lartillot, O.: Generalizability and simplicity as criteria in feature selection: Application to mood classification in music. IEEE Transactions in Audio, Language, and Speech Processing, 19 (6), 1802--1812 (2011)
15. Dibben, N.: The role of peripheral feedback in emotional experience with music. Music Perception, 22(1), 79--116 (2004)
16. Bigand, E., Vieillard, S., Madurell, F., Marozeau, J., & Dacquet, A.: Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts. Cognition and Emotion, 19(8), 1113--1139 (2005)
17. Ogg, M.: Physiological responses to music: measuring emotions. Undergraduate thesis. McGill University (2009)
18. Lartillot, O., & Toiviainen, P.: MIR in Matlab (II): A Toolbox for Musical Feature Extraction From Audio, Proceedings of the International Conference on Music Information Retrieval, Wien, Austria (2007)
19. Lartillot, O. Eerola, T., Toiviainen, P. Fornari, J.: Multi-feature modeling of pulse clarity: Design, validation, and optimization. In Proceedings of the International Symposium on Music Information Retrieval (2008)
20. Kim, J. and André, E.: Emotion Recognition Based on Physiological Changes in Listening Music, IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(12), 2067--2083 (2008)
21. Pan, J. and Tompkins, W.: A Real-Time QRS Detection Algorithm, IEEE Trans. Biomedical Eng., 32(3), 230—323 (1985)

CCA and a Multi-way Extension for Investigating Common Components between Audio, Lyrics and Tags.

Matt McVicar[1] and Tijl De Bie[2] *

Intelligent Systems Lab, University of Bristol
matt.mcvicar@bristol.ac.uk, tijl.debie@gmail.com

Abstract. In our previous work, we used canonical correlation analysis (CCA) to extract shared information between audio and lyrical features for a set of songs. There, we discovered that what audio and lyrics share can be largely captured by two components that coincide with the dimensions of the core affect space: valence and arousal. In the current paper, we extend this work significantly in three ways. Firstly, we exploit the availability of the Million Song Dataset with the MusiXmatch lyrics data to expand the data set size. Secondly, we now also include social tags from Last.fm in our analysis, using CCA also between the tag space and the lyrics representations as well as between the tag and the audio representations of a song. Thirdly, we demonstrate how a multi-way extension of CCA can be used to study these three datasets simultaneously in an incorporated experiment. We find that 2-way CCA generally (but not always) reveals certain mood aspects of the song, although the exact aspect varies depending on the pair of data types used. The 3-way CCA extension identifies components that are somewhere in between the 2-way results and, interestingly, appears to be less prone to overfitting.

Keywords: Canonical Correlation Analysis, Mood Detection, Million Song Dataset, MusiXmatch, Last.fm.

1 Introduction

In this paper we ask what is shared between the audio, lyrics and social tags of popular songs. We employ canonical correlation analysis (CCA) to find maximally correlated projections of these three feature domains in an attempt to discover commonalities and themes. In our previous work [16] we attempted to maximise the correlation between audio and lyrical features and discovered that the optimal correlations related strongly to the mood of the piece.

We extend this work significantly in three ways. Firstly, we make use of the recently-available Million Song Dataset (MSD,[1]) to gather a large number of audio and lyrical features, verifying our previous work on a larget dataset. Secondly, we incorporate a third feature space based on social tags from Last.fm[1].

* This work was partially supported by the EPSRC grant number EP/E501214/1

[1] www.last.fm

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

On these three datasets we are able to conduct pairwise 2-dimensional CCA on the largest public dataset of this type currently available. Lastly, we demonstrate how 3-dimensional CCA can be used to investigate these data simultaneously, leading to a multi-modal analysis of three aspects of music. Whilst it was intuitive to us in our previous work that lyrics and audio would have mood in common, it is less clear to us what commonalities are shared between the other pairs of datasets. We therefore take a more serendipitous approach in this study, aiming to discover which features are most strongly related.

The rest of this paper is arranged as follows. In the remainder of this Section we discuss the relevant literature and background to our work. We detail our data collection methods, feature extraction, and framework in Section 2. Section 3 deals with the theory of CCA in 2 and 3 dimensions. In Section 4 we present our findings, which are discussed and concluded in Section 5.

1.1 The Core Affect Space

Although it may be the case that our CCA analysis leads to components other than emotion, we suspect that many will relate to the mood of the piece. We therefore review the analysis of mood in this Subsection.

Russell [17] proposed a method for placing emotions onto a two-dimensional *valence-arousal* space, known in psychology as the *core affect space* [18]. The valence of a word describes its attractiveness/aversiveness, whilst the arousal relates to the strength, energy or activation. An example of a high valence, high arousal word is ecstatic, whilst depressed would score low on both valence and arousal. A third dimension describing the dominance of an emotion has also been suggested [6], but rarely used by researchers. A more detailed visualisation of the valence/arousal space with example words is shown in Figure 1.

1.2 Relevant Works

The valence/arousal space has been used extensively by researchers in the field of automatic mood detection from audio. Harmonic and spectral features were used by [8], whilst in [5] they utilised low-level features such as the spectral centroid, rolloff, flux, slope, skewness and kurtosis. Time-varying features in the audio domain were employed by various authors [15, 20], which included MFCCs and short time Fourier transforms. For classification, many authors have utilised SVMs, which have been shown to successfully discriminate between features [9].

In the lyrical domain, [7] used bag-of-words (BoW) models as well as n-grams and term frequency-inverse document frequency (TFIDF) to classify mood based on lyrics, whilst [10] made use of the experimentally deduced affective norms of english words (ANEW) to assign valence and arousal scores to individual words in lyrics. Both of these studies were conducted on sets of 500-2,000 songs.

The first evidence of combining text and audio in mood classification can be seen in [21]. They employed BoW text features and psychological features for classification and demonstrated a correlation between the verbal emotion features and the emotions experienced by the listeners on a small set of 145

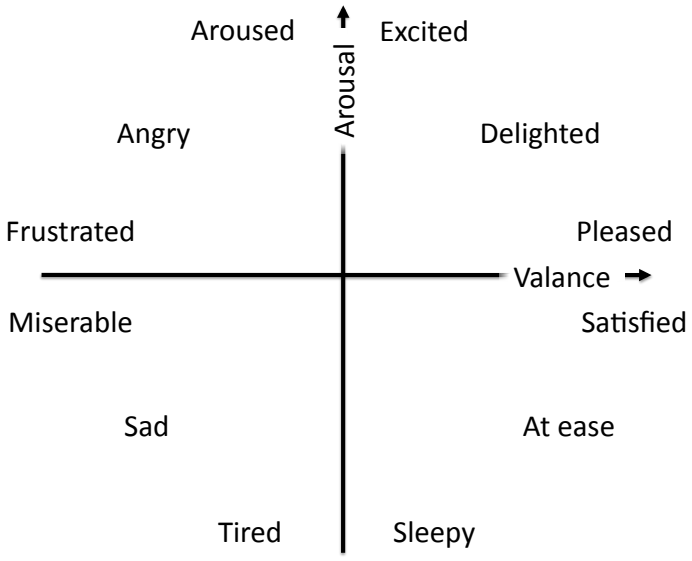


Fig. 1: The 2-dimensional valence/arousal space as proposed by Russell [17]. Words with high valence are more positive, whilst low valence words are pessimistic. High/low arousal words are energetic/restful respectively.

songs. A larger study was conducted in [13] where they classified 1,000 songs into 4 mood categories and found that by combining audio and lyrical features an increase in recognition accuracy was observed.

In the tag domain, [14] used the social website Last.fm to create a semantic mood space using latent semantic analysis. Via the use of a self-organising map, they reduce this high-dimensional space to a 2-D representation and compared this to Russell's valence/arousal space, with encouraging results.

In combining tag and audio data, [3] demonstrated that tag features were more informative than audio, whilst the combination was more informative still. This was conducted on a set of 1,612 songs and up to 5 mood or theme categories. Finally, a recent study considered regression of musical mood in continuous dimensional space using combinations of audio, lyrics and tags on a set of 2,648 UK pop songs [19].

Whilst insightful in terms of features and classification techniques, all of the studies previously mentioned were conducted on small datasets by todays standards (all significantly less than 10,000 songs). In this paper we address this issue in a truly large-scale, multi-modal analysis. We discuss our feature extraction and framework for our analysis in the following Section.

2 Data Collection & Framework

This section details our data collection methods and the motivation for our approach. We found the overlap of the Million Song, MusiXmatch and Last.fm datasets to be 223,815 songs in total, which was comprised of 197,436 training songs and 26,379 test songs. After removing songs which contained empty features, no lyrics or no tags, as well as those not in English, we were left with 101,235 (88%) training songs and 13,502 test songs (12%).

2.1 The Million Song Dataset

Devised as a way for researchers to conduct work on musical data without the need to purchase a large number of audio files, the Million Song Dataset was released on Feb 8^{th}, 2011. We downloaded this dataset in its entirety and extracted from it features relating to the audio qualities of the music. The features we specifically computed are shown in Table 1. We also give our interpretation of the features extracted, although there are some (e.g. danceability) where we are unsure of the feature extraction process.

Table 1: List of audio features extracted from the million song dataset, with interpretations.

Feature	Interpretation
Mean Bar Confidence	Average bar stability
Std Bar Confidence	Variation in bar stability
Mean Beat Confidence	Average beat stability
Std Beat Confidence	Variation in beat stability
Danceability	Danceability of track
Duration	Total track time in seconds
Key	Track harmonic centre (major keys only)
Key Confidence	Confidence in Key
Loudness	Loudness of track
Mode	Modality (major or minor) of track
Mode Confidence	Confidence in Mode
Mean Sections Confidence	Average confidence in section boundaries
Std Sections Confidence	Variation in section boundary confidences
Mean Seg. Conf.	Average confidence in segment boundaries
Mean Timbres 1-12	12 features relating to average sound quality
Std Timbres 1-12	12 features related to variation in sound quality
Tempo	Speed in Beats Per Minute
Loudness Max	Total maximum of track loudness
Loudness Start	Local max of loudness at the start of the track
Tatums Confidence	Confidence in tatum prediction
Time Signature	Predicted number of beats in a bar
Time Signature Confidence	Confidence in time signature

2.2 MusiXmatch

An addition to the MSD, the MusiXmatch dataset contains lyrical information for a subset of the million songs. The features are stored in bag-of-words format (for copyright reasons), and are stemmed versions of the top 5,000 words in the database. In order to ensure we had meaningful words, we restricted ourselves to the words which were part of the ANEW dataset [4], which reduced our dataset to 603 words. We converted the BoW data to a term frequency-inverse document frequency (TFIDF) score [11] via the following transformation.

Let the term frequency of the i^{th} feature from the j^{th} song be simply the BoW feature normalised by the count of this lyric's most frequent word:

$$TF_{i,j} = \frac{|\text{word } i \text{ appears in lyric } j|}{\text{maximum word count of lyric } j}$$

where $|\cdot|$ denotes 'number of'. The inverse document frequency measures the importance of a word in the database as a whole and is calculated as:

$$IDF_i = \log \frac{\text{total number of lyrics}}{|\text{lyrics containing word } i| + 1}$$

(we include the +1 term to avoid potentially dividing by 0). The TFIDF score is then the product of these two values:

$$TFIDF_{i,j} = TF_{i,j} \times IDFi$$

The TFIDF score gives an indication of the importance of a word within a particular song and the entire database. Note that we used the ANEW database simply to construct a dictionary of words which contain some emotive content - no experimental valence/arousal or mood scores were incorporated into our feature matrix.

2.3 Last.fm Data

The Last.fm data contains information on user-generated tags and artist similarities, although we neglect the latter for the purpose of this study. The dataset contains information on 943,347 tracks matched to the MSD and tag counts for each song. We discovered 522,366 unique tags although only considered tags which appeared in at least 1,000 songs, which resulted in 829 features. The top tags from the dataset were *Rock, Pop, Alternative, Indie* and *Electronic*. We constructed a TF-IDF score for each tag in each song analogously to the previous section. Although it would have been possible to filter the tags according to the ANEW database as per the lyrics, we know that tags contain information other than mood, such as genre data. We are optimistic that our algorithm may pick up such information, and so did not filter the Last.fm tags.

2.4 Framework

In our previous work [16] we introduced an exploratory framework for the use of CCA in correlating audio and lyrical features. We briefly recap this framework for 2-way CCA before extending it to use in 3 datasets.

We are interested in what is consistent between the audio, lyrics and tags of a song. In previous work, researchers have searched for a function f which maps audio to mood [$f(audio) = mood$], else from lyrics or tags [$g(lyrics) = mood$, $h(tags) = mood$]. In our 2-way CCA we seek functions which satisfy one of:

$$f(audio) \approx g(lyrics)$$
$$f(audio) \approx h(tags)$$
$$g(lyrics) \approx h(tags)$$

to a good approximation and for a large number of songs. Previously, we assumed that the first relationship in the above equations captured some aspect of mood, knowing of no other commonalities between the audio and lyrics of a song. This was verified by using 2-way CCA to find such functions f and g. In this study, we take a more serendipitous approach. We will use 2-way CCA on each pair of datasets and see which kinds of commonalities are found. Perhaps they will relate to mood, but we hope to discover other relationships and correlations within the data. The extension of this work to 3 dimensions follows a similar framework. We now seek functions f, g and h such that:

$$f(audio) \approx g(lyrics) \approx h(tags) \tag{1}$$

simultaneously. Again, these functions will not hold true for every song, but we hope they are approximately true for a large number of songs. The next Section deals with the theory of canonical correlation analysis.

3 Canonical Correlation Analysis and a 3-Way Extension

3.1 2-Way CCA

Given two datasets $X \in \mathbb{R}^{n \times d_x}$ and $Y \in \mathbb{R}^{n \times d_y}$, canonical correlation analysis finds what is consistent between them. This is realised by finding projections of X and Y through the dataset which maximise their correlation. In other words, we seek weight vectors $w_x \in \mathbb{R}^{d_x}$, $w_y \in \mathbb{R}^{d_y}$ such that the angle θ between Xw_x and Yw_y is minimised:

$$\{w_x^*, w_y^*\} = \underset{w_x, w_y}{\operatorname{argmin}} \ \theta(Xw_x, Yw_y)$$

Conveniently, this can be realised as a generalised eigenvector problem (a full derivation can be found in, for example, [2]):

$$\begin{pmatrix} 0 & X^TY \\ Y^TX & 0 \end{pmatrix} \begin{pmatrix} w_x \\ w_y \end{pmatrix} = \lambda \begin{pmatrix} X^TX & 0 \\ 0 & Y^TY \end{pmatrix} \begin{pmatrix} w_x \\ w_y \end{pmatrix} \tag{2}$$

In our experiments, X and Y will represent data matrices formed from the MSD, MusiXmatch or Last.fm datasets. The eigenvalue λ is the achieved correlation between the two datasets and (w_x, w_y) are the importance of each vector in the corresponding data space. The eigenvectors corresponding to λ can be sorted by magnitude to give a rank of feature importance in each of the data spaces.

3.2 3-Way CCA

Whilst it will be insightful to see the pairwise 2-way correlations between the three datasets, it would be more satisfying to investigate what is consistent between all 3 simultaneously. Various ways of exploring this have been explored in [12] - a natural extension in our setting can be motived as follows. Consider three datasets $X \in \mathbb{R}^{n \times d_x}$, $Y \in \mathbb{R}^{m \times d_y}$, $X \in \mathbb{R}^{p \times d_z}$. We motivate the correlation of these three variables graphically. Consider 3 datasets and (for ease of plotting) 3 songs within this set. A potential set of projections Xw_X, Yw_Y, and Zw_Z is shown in Figure 2.

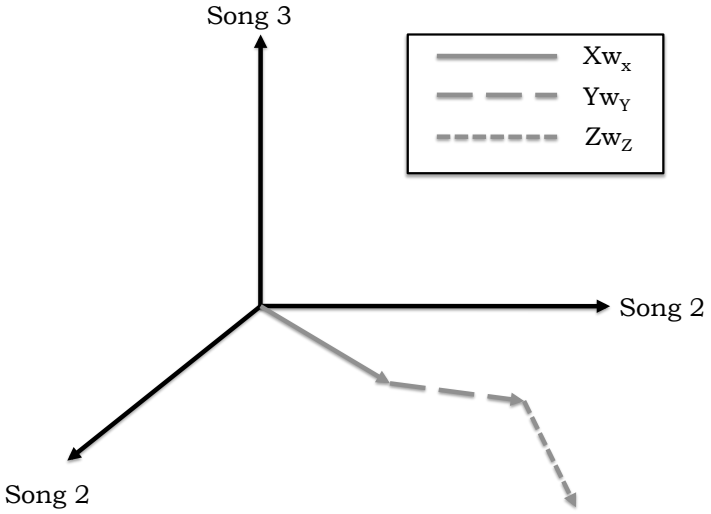


Fig. 2: Motivation for 3-way CCA on 3 example songs, showing the projections Xw_X, Yw_Y, Zw_Z.

It is clear that the three projections are strongly correlated if the norm of their sum is large. However, this is easy to obtain if each of the projections is arbitrarily large. We therefore enforce the constraint that the individual lengths

are bounded, and solve the following optimization problem:

$$\max_{w_x,w_y,w_z} \|Xw_x + Yw_y + Zw_z + 1\|^2$$
$$s.t. \quad \|Xw_X\|^2 + \|Yw_Y\|^2 + \|Zw_Z\|^2 = 1$$

Solving the above via the method of Lagrange multipliers, we obtain

$$\frac{1}{2}\frac{\partial}{\partial w_*}\Big[\|Xw_X + Yw_Y + Zw_Z\|^2 - \lambda\Big(\|Xw_X\|^2 + \|Yw_Y\|^2 + \|Zw_Z\|^2\Big)\Big] = 0$$

where the asterix $*$ represents partial differentiation with respect to the appropriate variable. This leads to the simultaneous equations

$$X^TXw_X + X^TYw_Y + X^TZw_Z - \lambda X^TXw_X = 0$$
$$Y^TXw_X + Y^TYw_Y + Y^TZw_Z - \lambda Y^TXw_Y = 0$$
$$Z^TXw_X + Z^TYw_Y + Z^TZw_Z - \lambda Z^TZw_Z = 0$$

which, in matrix form, is

$$\begin{pmatrix} 0 & X^TY & X^TZ \\ Y^TX & 0 & Y^TZ \\ Z^TX & Z^TY & 0 \end{pmatrix} \begin{pmatrix} w_X \\ w_Y \\ w_Z \end{pmatrix} = (\lambda - 1) \begin{pmatrix} X^TX & 0 & 0 \\ 0 & Y^TY & 0 \\ 0 & 0 & Z^TZ \end{pmatrix} \begin{pmatrix} w_X \\ w_Y \\ w_Z \end{pmatrix} \quad (3)$$

Substituting $\lambda \to \lambda - 1$, we see that 3-dimensional CCA is an obvious extension of the 2-dimensional set-up seen in Equation 2. Note however that the λ is now a generalisation of the notion of correlation, and is not necessarily bounded in absolute value by 1. In our setting, the datasets X, Y and Z represent the MSD, MusiXmatch and Last.fm datasets and our aim will be to maximise the correlation between them. Our experimental results using pairwise CCA and 3-way CCA are presented in the next Section.

4 Experiments

4.1 Audio - Lyrical CCA

We begin with a reproduction of our previous work [16] which uses CCA on audio and lyrical datasets. This will serve to verify our method scales to datasets of realistic sizes. The projections of the Audio and Lyrical datasets, ranked by test correlation magnitude, are shown in Table 2. In each pairwise CCA experiments we found the significance of the correlations under a χ^2 distribution to be numerically 0, owing to the extremely large data sizes. It is therefore more important to look at the magnitude of the correlations rather than their significance in the following experiments.

These projections agree with our previous finding that mood is one of the common components between audio and lyrics. In the first component, words

Table 2: Features with largest weights using Audio and Lyrical features in 2-way CCA, first 3 CCA components. Training correlations on the first three components were 0.5032, 0.4484 and 0.2409 whilst the corresponding test correlations were 0.5034, 0.4286 and 0.2875.

CCA Comp.	**Lowest**		**Highest**	
	Lyrical Feature	**Lyrical Weight**	**Lyrical Paper**	**Lyrical Weight**
1	Death	-0.0358	Love	0.0573
	Dead	-0.0274	Baby	0.0394
	Burn	-0.0239	Blue	0.0197
	Hate	-0.0219	Girl	0.0190
	Pain	-0.0204	Man	0.0170
1	**Audio Feature**	**Audio Weight**	**Audio Feature**	**Audio Weight**
	Loudness Max	-0.6824	Mean Timbre 1	0.6559
	Loudness	-0.0711	Mean Seg. Conf.	0.1638
	Duration	-0.0413	Loudness Start	0.1539
	Mean Timbre 10	-0.0311	Mean Timbre 5	0.0698
	Std Timbre 6	-0.0222	Mean Timbre 6	0.0649
	Lowest		**Highest**	
	Lyrical Feature	**Lyrical Weight**	**Lyrical Feature**	**Lyrical Weight**
2	Dream	-0.0182	Man	0.0354
	Love	-0.0177	Hit	0.0325
	Heart	-0.0142	Girl	0.0303
	Fall	-0.0117	Rock	0.0291
	Lonely	-0.0113	Baby	0.0268
2	**Audio Feature**	**Audio Weight**	**Audio Feature**	**Audio Weight**
	Loudness Max	-0.5568	Mean Timbre 1	0.7141
	Loudness Start	-0.2846	Loudness	0.1424
	Std Seg. Conf.	-0.0855	Std Timbre 8	0.1233
	Std Timbre 4	-0.0525	Mean Seg. Conf.	0.1227
	Std Timbre 1	-0.0402	Mean Timbre 8	0.0446
	Lowest		**Highest**	
	Lyrical Feature	**Lyrical Weight**	**Lyrical Feature**	**Lyrical Weight**
3	Baby	-0.0304	Man	0.0572
	Fight	-0.0281	Love	0.0409
	Hate	-0.0223	Dream	0.0341
	Girl	-0.0223	Child	0.0301
	Scream	-0.0199	Dark	0.0295
3	**Audio Feature**	**Audio Weight**	**Audio Feature**	**Audio Weight**
	Mean Timbre 1	-0.6501	Loudness Max	0.5613
	Loudness Start	-0.2281	Duration	0.1874
	Std Timbre 6	-0.1507	Loudness	0.1377
	Std Seg. Conf.	-0.0898	Std Timbre 8	0.1050
	Tatums Conf.	-0.0850	Std Timbre 10	0.0891

with low weights appear more aggressive, whilst more optimistic words have the highest weights. This suggests that this CCA component has captured the notion of valence. Audio features in this domain show that high valence songs are loud, whilst low valence words have important timbre features.

The second CCA component appears to have identified relaxed lyrics at one extreme and more active words at the other. We consider this to be a realisation of the arousal dimension. In the audio domain, loudness and timbre again seems to play an important role. It is more difficult to interpret the third CCA component, although the sharp decay of test correlation values show that the first two CCA components dominate the analysis.

4.2 Audio - Tag CCA

We now investigate 2-way CCA on audio/tag data, using Last.fm tags in place of the lyrical data from Subsection 4.1. Components 1-3 are shown in Figure 3.

The first component of this CCA analysis seems to have found that the maximal correlation can be obtained by having tags associated with metal tags at one extreme and more serene tags at the other. The audio features in this CCA component seems to be well described by the later timbre features.

In the second component, we also see an obvious trend, with modern urban genre tags receiving high weights and more traditional music at the other. In the audio space, these genres seem to be associated with timbre and audio features.

The correlations between these two sets is so strong that we can even interpret the third CCA component, which has identified modern electronic music and acoustic blues/country as strongly opposing tags in this dimension. Interestingly, components 2 and 3 appear to have identified two distinct types of 'oldies' music (folk/blues respectively). In the audio domain these are accompanied by structural stability (segment/tatum confidence) features.

4.3 Lyrical - Tag CCA

The first three CCA components of this experiment are shown in Figure 4.

In the first component it seems we are distinguishing heavy metal genres from less aggressive styles. In the lyrical domain we see that the words with low weights hold strongly negative valence; those with high weights are more optimistic. The authors find the notion of Melodic Black Metal somewhat oxymoronic.

The second component also has a clear trend - extremes in this dimension appear to be hip-hop/rap vs. worship music. We postulate that this represents the dominance dimension mentioned in the Introduction, with the lyrical weights corroborating this. In the third component we see no particular trend, which is supported by the low correlation of 0.1807. Comparison with the training correlation of 0.4826 suggests that this component is suffering from overfitting.

4.4 3-way Experiment

We display our results from 3-way CCA in Table 5.

Table 3: Features with largest weights using Audio and Tag Features in 2-way CCA, first 3 CCA components. Training correlations on the on these components were 0.7361, 0.6432 and 0.5725 whilst the corresponding test correlations were 0.5685, 0.5237 and 0.3428 respectively.

CCA comp.	**Lowest**		**Highest**	
	Tag Feature	**Tag Weight**	**Tag Feature**	**Tag Weight**
1	Female Vocalists	-0.0352	Metal	0.0672
	Acoustic	-0.0304	Death Metal	0.0542
	Singer-Songwriter	-0.0289	Brutal Death Metal	0.0425
	Classic country	-0.0271	Punk rock	0.0378
	Folk	-0.0265	Metalcore	0.0371
	Audio Feature	**Audio Weight**	**Audio Feature**	**Audio Weight**
1	Mean Timbre 1	-0.5314	Loudness Max	0.7460
	Loudness Start	-0.1700	Std Timbre 6	0.0988
	Mean Timbre 6	-0.1558	Mean Timbre 2	0.0500
	Mean Seg. Conf.	-0.1469	Mean Timbre 3	0.0491
	Mean Timbre 5	-0.1021	Std bar Conf.	0.0267
	Lowest		**Highest**	
	Tag Feature	**Tag Weight**	**Tag Feature**	**Tag Weight**
2	Oldies	-0.0153	Hip-Hop	0.0418
	Beautiful	-0.0132	Dance	0.0355
	60s	-0.0126	Hip hop	0.0353
	Singer-Songwriter	-0.0116	Rap	0.0351
	Folk	-0.0110	Rnb	0.0231
2	**Audio Feature**	**Audio Weight**	**Audio Feature**	**Audio Weight**
	Loudness Start	-0.5069	Mean Timbre 1	0.7522
	Loudness Max	-0.3506	Loudness	0.1248
	Mean Timbre 6	-0.0631	Std Timbre 8	0.0578
	Std Timbre 1	-0.0374	Mean Timbre 4	0.0497
	Std Seg. Conf.	-0.0360	Mean Timbre 10	0.0415
	Lowest		**Highest**	
	Tag Feature	**Tag Weight**	**Tag Feature**	**Tag Weight**
3	Electronic	-0.0284	Oldies	0.0335
	Dance	-0.0220	Classic Blues	0.0325
	Vocal Trance	-0.0198	Classic country	0.0290
	Epic	-0.0186	50s	0.0279
	Pop	-0.0181	Delta blues	0.0250
	Audio Feature	**Audio Weight**	**Audio Feature**	**Audio Weight**
3	Mean Timbre 1	-0.6988	Loudness Max	0.6416
	Mean Timbre 4	-0.1141	Mean Timbre 3	0.1404
	Tatums Conf.	-0.0649	Mean Seg. Conf.	0.0757
	Duration	-0.0589	Mean Timbre 6	0.0732
	Std Segs Conf.	-0.0556	Loudness Start	0.0507

Table 4: Features with largest weights using Lyrical and Tag Features in 2-way CCA, first three CCA components. Training correlations on these components were 0.5828, 0.4990 and 0.4826 whilst test correlations were 0.3984, 0.3713 and 0.1807 respectively.

CCA comp.	**Lowest**		**Highest**	
	Lyrical Feature	**Lyrical Weight**	**Lyrical Feature**	**Lyrical Weight**
1	Death	-0.1851	Love	0.2330
	Dead	-0.1201	Baby	0.1807
	Human	-0.1049	Girl	0.0792
	God	-0.0993	Christmas	0.0726
	Pain	-0.0925	Blue	0.0679
	Tag Feature	**Tag Weight**	**Tag Feature**	**Tag Weight**
1	Brutal Death Metal	-0.3029	Xmas	0.0785
	Death Metal	-0.2470	Female Vocalists	0.0718
	Metal	-0.2449	Oldies	0.0688
	Melodic black metal	-0.2029	Pop	0.0680
	Black metal	-0.1338	Rnb	0.0652
	Lowest		**Highest**	
	Lyrical Feature	**Lyrical Weight**	**Lyrical Feature**	**Lyrical Weight**
2	Hit	-0.1448	Christmas	0.4082
	Man	-0.1267	Snow	0.0907
	Rock	-0.1180	Glory	0.0607
	Money	-0.1073	Joy	0.0549
	Brother	-0.0999	Angel	0.0530
	Tag Feature	**Tag Weight**	**Tag Feature**	**Tag Weight**
2	Hip hop	-0.2312	Xmas	0.4111
	Rap	-0.2014	Christmas	0.1679
	Hip-Hop	-0.1927	Christian	0.0662
	Gangsta Rap	-0.1460	Female Vocalists	0.0501
	Underground hip hop	-0.1143	Worship	0.0480
	Lowest		**Highest**	
	Lyrical Feature	**Lyrical Weight**	**Lyrical Feature**	**Lyrical Weight**
3	Love	-0.0273	Christmas	0.6031
	Heart	-0.0262	Snow	0.0992
	Rain	-0.0247	Man	0.0800
	Alone	-0.0229	Rock	0.0716
	Dream	-0.0224	Hit	0.0702
	Tag Feature	**Tag Weight**	**Tag Feature**	**Tag Weight**
3	Love	-0.0399	Xmas	0.5932
	Female vocalists	-0.0257	Christmas	0.2381
	Alternative rock	-0.0252	Hip hop	0.1265
	Rain	-0.0237	Rap	0.0975
	Oldies	-0.0227	Hip-Hop	0.0906

Table 5: Summary of 3-way CCA analysis. CCA components are shown in rows, with the highest and lowest-weighted features of each data space (audio, lyrical, tag) occupying the columns. The generalised training correlations on the first three components were found to be 2.1749, 2.0005, and 1.76559 whilst the generalised test correlations were found to be 2.1809,2.0036 and 1.7595 (recall that these generalised correlations are not necessarily bounded in absolute value by 1). Abbreviations: DM = Death Metal, BM = Black Metal, SS = Singer-Songwriter, FV = Female Vocalists, AR = Alternative Rock, UHH = Underground hip hop.

CCA	**Lowest**		**Highest**		**Lowest**		**Highest**		**Lowest**		**Highest**	
Comp.	Audio Feature	Weight	Audio Feature	Weight	Word	Weight	Word	Weight	Tag Feature	Weight	Tag Feature	Weight
1	Loudness Max	-0.7008	Mean Timbre 1	0.6064	Death	-0.0346	Love	0.0572	Metal	-0.0581	FVs	0.0242
	Loudness	-0.0442	Loudness Start	0.1906	Dead	-0.0272	Baby	0.0382	DM	-0.0517	Pop	0.0189
	Std Timbre 6	-0.0426	Mean Segs Conf.	0.1553	Hate	-0.0220	Blue	0.0179	Brutal DM	-0.0510	Classic Country	0.0183
	Duration	-0.0305	Mean Timbre 6	0.0925	Burn	-0.0216	Girl	0.0171	Melodic BM	-0.0348	Oldies	0.0176
	Mean Timbre 10	-0.0255	Mean Timbre 5	0.0766	Pain	-0.0191	People	0.0147	Metalcore	-0.0267	Soul	0.0167
2	Loudness Max	-0.4676	Mean Timbre 1	0.6797	Dream	-0.0161	Hit	0.0372	Beautiful	-0.0183	Hip Hop	0.0630
	Loudness Start	-0.4107	Loudness	0.2063	Love	-0.0131	Man	0.0343	FVs	-0.0130	Hip-Hop	0.0625
	Std Segs Conf.	-0.0899	Std Timbre 8	0.1238	Heart	-0.0123	Rock	0.0320	Ambient	-0.0113	Rap	0.0596
	Std Timbre 1	-0.0568	Mean Segs Conf.	0.1031	Home	-0.0114	Girl	0.0291	Christian	-0.0112	Gangsta Rap	0.0333
	Std Timbre 4	-0.0469	Mean Timbre 10	0.0480	Sad	-0.0111	Baby	0.0283	Mellow	-0.0106	UHH	0.0260
3	Mean Timbre 1	-0.7004	Loudness Max	0.6402	Girl	-0.0124	Man	0.0268	Dance	-0.0266	Folk	0.0239
	Loudness Start	-0.1666	Loudness	0.0756	Fight	-0.0124	Blue	0.0177	Rock	-0.0207	Brutal DM	0.0208
	Mean Timbre 4	-0.0797	Mean Timbre 6	0.0672	Crash	-0.0107	Death	0.0176	Pop	-0.0183	Melodic BM	0.0195
	Std Timbre 6	-0.0683	Duration	0.0558	Alive	-0.0104	Christmas	0.0156	AR	-0.0174	Acoustic	0.0176
	Std Segs Conf.	-0.0547	Mean Timbre 3	0.0472	Baby	-0.0100	Dark	0.0136	Electronic	-0.0162	Classic Country	0.0171

In this incorporated experiment, the most prevalent dimension appears to relate to arousal - highly weighted tags and features are gentle in nature, with aggressive tags, lyrics and audio features. The second component seems to represent arousal. We struggle to find an explanation for the third component.

5 Discussion & Conclusions

In this Section, we discuss some of the findings from the previous Section, summarise the conclusions of our study and suggest areas for future work.

5.1 Discussion

It is clear there are similar components in this study across different experiments. For instance, the first component of the audio/lyrical 2-way CCA experiment in the lyrical domain (first few rows of Table 2) were very similar to the first component in the lyrical domain in the 3-way experiment (first rows of Table 5, second cell). It appears that both of these discovered dimensions are capturing the valence of the lyrics. To verify that these projections were indeed similar, we computed the correlation between them (ie Yw_Y from Table 2 with Yw_Y from Table 5), and found it to be 0.9979. The conclusion to be drawn is that the valence of lyrics is very easily captured, by comparing with audio and/or tag information.

We now turn our attention to the second CCA component. Interested in what 3-Way CCA analysis offered over pairwise CCA experiments, we investigated the correlations between each pair of lyrical and tag projections from all three experimental set-ups (2 pairwise and 3-Way). These are shown in Table 6.

Table 6: Comparison of Lyrical and Tag projections in pairwise and 3-way experiments.

(a) Lyrical Projections

CCA comp. 2	$\mathbf{YW_Y}$ Lyrics/Tags 3-Way CCA	
Lyrics/Audio	0.8679	0.9899
Lyrics/Tags	-	0.8886

(b) Tag Projections

CCA comp. 2	$\mathbf{ZW_Z}$ Tags/Lyrics 3-Way CCA	
Tags/Audio	0.7534	0.8853
Tags/Lyrics	-	0.9434

The first of these tables can be interpreted as follows. Recall that in the lyrics-audio CCA experiment we found the second component to describe the arousal of the lyrics. In the lyrics-tag space we found the second lyrical component related to the dominance of the lyrics. Recall that the correlations are equivalent to the angles between the projected datasets. Table 6(a) therefore shows that the cosines of the angles between these vectors and the third CCA component are 0.9899 and 0.8886 respectively, but that the cosine of the angle between

themselves is 0.8679. This shows that the 3-Way CCA component sits somewhere between arousal and dominance, which can be verified by looking at the top and bottom-ranked words in Tables 2, 3 and 5.

A similar, and in fact stronger pattern can be observed in tag space by investigating Table 6(b). Again, the 3-way CCA analysis seems to be an intermediate between the 'old vs new' dimension observed in the audio-tag space (Table 3, second component) and the dominance discovered in the lyrical-tag space (Table 4, second component).

5.2 Conclusions & Further Work

In this paper, we have conducted a large-scale study of the correlations between audio, lyrical and tag features based on the Million Song Dataset. By the use of pairwise 2-dimensional CCA we demonstrated that the optimal correlations between these datasets appear to have reconstructed the valence/arousal/dominance dimensions of the core affect space, even though this was in no way imposed by the algorithm. In some cases, we discovered components which appeared to capture some genre information, such as the third component of Table 3.

By using 3-dimensional CCA, we studied the 3 datasets simultaneously and discovered that valence and arousal were the most correlated features. The correlations beyond 2 or 3 components are difficult to interpret, which fits well studies which describe the core affect space as a 2 or 3 dimensional space.

In our future work we would like to investigate different multiway CCA extensions such as those seen in [12], perhaps on new datasets as they are released. We also would like to more thoroughly investigate regularization techniques to avoid overfitting.

References

1. Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.
2. T. De Bie, N. Cristianini, and R. Rosipal. Eigenproblems in pattern recognition. In E. Bayro-Corrochano, editor, *Handbook of Computational Geometry for Pattern Recognition, Computer Vision, Neurocomputing and Robotics*. Springer-Verlag, 2004.
3. K. Bischoff, C.S. Firan, R. Paiu, W. Nejdl, C. Laurier, and M. Sordo. Music mood and theme classification-a hybrid approach. In *Proc. of the Intl. Society for Music Information Retrieval Conf., Kobe, Japan*, 2009.
4. M.M. Bradley and P.J. Lang. Affective norms for english words (anew): Instruction manual and affective ratings. *University of Florida: The Center for Research in Psychophysiology*, 1999.
5. J.J. Burred, M. Ramona, F. Cornu, and G. Peeters. Mirex-2010 single-label and multi-label classification tasks: ircamclassification09 submission. *MIREX 2010*, 2010.

6. R. Cowie, E. Douglas-Cowie, B. Apolloni, J. Taylor, A. Romano, and W. Fellenz. What a neural net needs to know about emotion words. *Computational intelligence and applications*, pages 109–114, 1999.
7. H. He, J. Jin, Y. Xiong, B. Chen, W. Sun, and L. Zhao. Language feature mining for music emotion classification via supervised learning from lyrics. *Advances in Computation and Intelligence*, pages 426–435, 2008.
8. X. Hu and J.S. Downie. When lyrics outperform audio for music mood classification: a feature analysis. In *Proceedings of ISMIR*, pages 1–6, 2010.
9. X. Hu, J.S. Downie, C. Laurier, M. Bay, and A.F. Ehmann. The 2007 mirex audio mood classification task: Lessons learned. In *Proceedings of the 9th International Conference on Music Information Retrieval*, pages 462–467. Citeseer, 2008.
10. Y. Hu, X. Chen, and D. Yang. Lyric-based song emotion detection with affective lexicon and fuzzy clustering method. In *Proceedings of ISMIR*, 2009.
11. K.S. Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.
12. J.R. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451, 1971.
13. C. Laurier, J. Grivolla, and P. Herrera. Multimodal music mood classification using audio and lyrics. In *Machine Learning and Applications, 2008. ICMLA'08. Seventh International Conference on*, pages 688–693. IEEE, 2008.
14. C. Laurier, M. Sordo, J. Serra, and P. Herrera. Music mood representations from social tags. In *Proceedings of the 10th International Society for Music Information Conference, Kobe, Japan*. Citeseer, 2009.
15. M.I. Mandel. Svm-based audio classification, tagging, and similarity submissions. *online Proc. of the 7th Annual Music Information Retrieval Evaluation eX-change (MIREX-2010)*, 2010.
16. M. McVicar, T. Freeman, and T. De Bie. Mining the correlation between lyrical and audio features and the emergence of mood. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.
17. J.A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.
18. J.A. Russell. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145, 2003.
19. B. Schuller, F. Weninger, and J. Dorfner. Multi-modal non-prototypical music mood analysis in continous space: reliability and performances. In *Proc. of the 12th International Society for Music Information Retrieval (ISMIR) Conference*, pages 759–764, 2011.
20. K. Seyerlehner, M. Schedl, T. Pohle, and P. Knees. Using block-level features for genre classification, tag classification and music similarity estimation. *Submission to Audio Music Similarity and Retrieval Task of MIREX 2010*, 2010.
21. D. Yang and W.S. Lee. Disambiguating music emotion using software agents. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR04)*, pages 52–58, 2004.

Poster session 1:

Music Emotion: Analysis, Retrieval, and Multimodal Approaches, Synthesis, Symbolic Music-IR, Spatial Audio, Performance, Semantic Web

Music Emotion Regression based on Multi-modal Features[1]

Di Guan[1] , Xiaoou Chen[2] and Deshun Yang[3],
[1,2,3] Peking University
Institute of Computer Science & Technology
guandiyiyi417@gmail.com
{chenxiaoou, yangdeshun}@pku.edu.cn

Abstract. Music emotion regression is considered more appropriate than classification for music emotion retrieval, since it resolves some of the ambiguities of emotion classes. In this paper, we propose an AdaBoost-based approach for music emotion regression, in which emotion is represented in PAD model and multi-modal features are employed, including audio, MIDI and lyric features. We first demonstrate the effectiveness of our approach, and then focus on exploring the contribution of individual modalities to the regression of each emotion dimension. A series of experiments show that lyric contributes the most to the regression of emotion dimension P, while audio and MIDI contribute more to the regression of dimension A and D. Thinking that the three modalities provide complementary information from different angles, we combine them and show that the best regression performance is obtained when all modalities are used.

Keywords: Music emotion regression, Multi-modal, AdaBoost, PAD.

1 Introduction and Related Works

It is natural for us to organize and search music by emotional contents. Music emotion retrieval has gained increasing attention in the field of music information retrieval during the past few years [1].

Music emotion classification, in which the emotion space is modeled by a given number of classes, is a plausible approach to music emotion retrieval, but the emotional states within each class may vary a lot, and this ambiguity may confuse users when they retrieve music according to emotion. However in music emotion regression, the emotion space is viewed as continuous and each point in the space is considered as a distinctive emotional state [2]. In this way, the ambiguity associated with emotion classes can be successfully avoided, so music emotion regression is considered more appropriate for music emotion retrieval [3]. A regression approach is proposed for music emotion recognition in [3], the best performance evaluated in terms of the R^2 statistics reaches 58.3% for *arousal* and 28.1% for *valence*.

[1] Project supported by the Natural Science Foundation of China (Multi-modal Music Emotion Recognition technology research No.61170167) & Beijing Natural Science Foundation (Multimodal Chinese song emotion recognition)

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

In our work, PAD(Pleasure-Arousal-Dominance) emotion-state model is used to represent music emotion [4].Three nearly independent dimensions, P(pleasure), A(arousal) and D(dominance), are used to represent emotional states in PAD model. P distinguishes the positive-negative quality of emotional states, A refers to the intensity of physical activity and mental alertness, and D is defined in terms of control versus lack of control. In our work, we normalize all dimensions in the range of -4 to 4, in this way each emotional state corresponds to a specific point in PAD model. For example, "anger" corresponds to (-3.51, 2.59, 0.95), which indicates that it is a highly unpleasant, highly aroused, and moderately dominant emotional state.

Audio features have been commonly used in music emotion recognition and audio-based techniques could achieve promising results [5]. As a complementary source, lyric contains rich semantic information of songs and more emotionally relevant information which is not included in audio [6].Additionally, MIDI is used in symbolic music information retrieval [7]. Some previous works applied multi-modal features for music emotion recognition and achieved promising performance [8,9,10].

In this paper, we present an AdaBoost approach for music emotion regression where three-modality features, audio, MIDI and lyric, are employed. We firstly demonstrate the effectiveness of our regression approach by comparing it with several baseline regression algorithms, and secondly use each modality alone to explore the contribution of each modality to the regression of different emotion dimension, and thirdly combine the three modalities to demonstrate the performance improvement of multi-modal feature combination for music emotion regression, lastly use a feature selection technique to reduce feature dimensions and computational complexity.

The paper is organized as follows. Section 2 describes the features and feature processing, Section 3 describes the regression approach, Section 4 provides the analysis of experiment results, and the last Section makes the conclusion and prospect.

2 Dataset and Feature Processing

2.1 Dataset

We download 2500 Chinese songs from network, including audio, MIDI and lyric data for each song, which cover more than 900 singers and more than 1000 albums, and include different genres such as pop, rap and rock. Then 11 volunteers whose ages are from 22 to 50 use Self Assessment Manikins(SAM) [11] to annotate the songs with PAD values ranging from -4 to 4. When a song is annotated by more than 8 volunteers and the emotion values given by different annotators are consistent (all positive or all negative), the song will get a mean value as its emotion label and be retained into our dataset. In this way, the final music dataset includes 1687 songs.

2.2 Features

Audio Features. We extract audio features from wave files of the dataset by jAudio [12], which is a system to extract the basic features from audio signal. We set the window size to 512ms (the signal sampling rate to 22KHz) to extract audio features,

including one-dimension (e.g. RMS) and multi-dimension vectors (e.g. MFCC's). 27 kinds of audio features that have been commonly used in MIR are extracted to compose an audio feature vector of 112 dimensions for a song. Table 1 shows part of the audio features.

MIDI Features. We extract MIDI features from MIDI files of the dataset by jSymbolic [13], which is a feature extraction system for extracting high-level musical features from symbolic music representations, specifically from MIDI files. Unlike audio data, MIDI data contains the information reflecting music concepts directly. 102 kinds of MIDI features are extracted to compose a MIDI feature vector of 1022 dimensions for a song. Table 2 shows part of the MIDI features.

Table 1. The partial list of audiofeatures.

Audio features	
Feature	**Dimensions**
MFCC's	13
LPC	10
Spectral Rolloff	1
Spectral Flux	1
RMS	1
Compactness	1
Zero Crossings	1
…	…
Power Spectrum	variable
All	**112**

Table 2. The partial list of MIDI features.

MIDI features	
Feature	**Dimensions**
Basic Pitch Histogram	128
Beat Histogram	161
Melodic Interval Histogram	128
Pitch Class Distribution	12
Acoustic Guitar Fraction	1
Amount of Arpeggiation	1
Note Density	1
…	…
Duration	1
All	**1022**

Lyric Features. We firstly download the lyrics of all the songs from Internet, and then do some pre-processing to them with traditional NLP tools, including stop-words filtering and word segmentation etc. Finally Unigram, Bigram and Trigram features are extracted from the lyrics.

Unigram. Unigram refers to the sequences of single word appeared in documents.

Bigram. Bigram refers to a distinctive term containing 2 consecutive words appeared in documents. Because negation words often reverse emotion of the words next to them, it seems reasonable to incorporate word-pairs to take effect of negation words into account in emotion analysis.

Trigram. Trigram refers to a distinctive term containing 3 consecutive words appeared in documents. Because bigrams only reflect parts of useful multi-word patterns for emotion expression, we take trigrams into account additionally.

Finally, in order to reduce the lyric feature space, we select the 3000 most frequently appeared N-grams (*n=1, 2, 3*) as lyric features. In our work, the feature vector of a lyric can be expressed as (v_1, v_2, v_3, ... v_{3000}). Here $v_i \in \{0, 1\}$: if N-gram *i* appeared in the lyric, v_i=1; otherwise v_i=0.

2.3 Feature Processing

In our work, seven different feature sets are employed for the regression of emotion dimension P, A and D, including the set of audio features(A), the set of MIDI

features(M), the set of lyric features(L), the set of audio and MIDI features(A+M), the set of audio and lyric features(A+L), the set of MIDI and lyric features(M+L), and the set of audio, MIDI and lyric features(A+M+L). A simple concatenation scheme is employed to combine the multi-modal features. For example, a concatenated feature vector of the three modalities can be expressed as (A_1,A_2,...A_x , M_1,M_2,...M_y , L_1,L_2,...L_z). Where A_1~A_x are audio features, and $x=112$; M_1~M_y are MIDI features, and $y=1022$; L_1~L_z are lyric features, and $z=3000$. The number of dimensions of a concatenated feature vector is $x+y+z=4134$.

The space formed by the raw concatenated features has a huge number of dimensions. To reduce the computational complexity of learning and regression, increase the efficiency and generalization capability of the regression model, we do feature selection on each of the original feature sets, to find a subset of the original set which could maximize the performance of regression model. Correlation-based Feature Subset Selection [14] with BestFirst as its search method is employed in our work, which evaluates the worth of a feature subset by considering the individual predictive ability of each feature along with the degree of redundancy between them, subsets of features that are highly correlated with the class while having low inter-correlation are preferred [15].

In feature selection process, we have found that some features are effective to all the 3 emotion dimensions, such as LPC, beat histogram, basic pitch histogram, melodic interval histogram, etc. But some features only effective to some of the 3 dimensions, such as staccato incidence, spectral rolloff point, etc, which only effective to dimension A and D. After feature processing, we get seven final feature sets for emotion dimension P, A and D.Table 3 shows the number of selected features of each feature set.

Table 3. The number of selected features in each feature set.

	P	**A**	**D**
A	17	16	16
M	44	43	54
L	262	410	474
A+M	43	51	54
A+L	349	208	331
M+L	115	67	203
A+M+L	116	69	86

3 Regression Algorithm

AdaBoost is a commonly used boosting method, which works by iteratively running weak learners on different distributions of training data, so as to get an integrated regression model more powerful than weak learners.

We present an AdaBoost regression approach in this paper, which follows most of the steps of AdaBoost.R2 [16] and uses MultiLayerPerceptron(MLP) [17] as the weak learner. MLP is a typical feed forward neural network connecting several perceptrons by a hierarchy, and uses error back propagation to adjust connection weights

continuously. We called our approach AdaBoost.RM(R refers to Regression and M-MultiLayerPerceptron).

Given a set of m training instances: $(x_1, y_1),\ldots,(x_m, y_m)$, where $x_1 \ldots x_m$ are the features, and $y_1 \ldots y_m \in [-4, +4]$ are the P, A or D emotion values of the instances. Initially, we set the weights of the training instances as $D_t(i) = 1/m$, then iteratively running MLP on the instances to train a regression model for P, A, or D and modify the weights of the instances accordingly. We set the number of iterations to 10, because the performance of the algorithm no longer improves when the number of iterations is greater than 10. The instance weight modification method is as follows:

$$\overline{L_t} = \sum_{i=1}^{m} \left(\frac{f_t(x_i) - y_i}{\max_{i=1,2\ldots m}(f_t(x_i) - y_i)} \right)^2 D_t(i) \tag{1}$$

$$D_{t+1}(i) = \frac{D_t(i) \left(\frac{\overline{L_t}}{1 - \overline{L_t}} \right)^{(1 - L_t(i))}}{Z_t} \tag{2}$$

Where f_t is the regression model learned in iteration t, $f_t(x_i)$ is the regression result of x_i from model f_t. $D_t(i)$ is the weight of instance i in iteration t, $\overline{L_t}$ is the average loss of f_t, z_t is a normalization factor that makes $\sum_i D_{t+1}(i) = 1$.

This reweighting procedure makes the poorly predicted instances get higher weights but well predicted ones get lower weights. Finally, an average formula is used to calculate the final regression result instead of the "*INF*" formula of AdaBoost.R2:

$$f_{final}(x) = AVE \left[y \in Y: \sum_{t: f_t(x) \le y} \log \frac{1 - \overline{L_t}}{\overline{L_t}} \ge \frac{1}{2} \sum_t \log \frac{1 - \overline{L_t}}{\overline{L_t}} \right] \tag{3}$$

Where $Y = \{f_1(x), f_2(x), \ldots, f_T(x)\}$, *AVE* is the average function.

4 Experiments and Results Analysis

4.1 Evaluation Criteria

We conduct a series of experiments to evaluate the performance of our regression approach. Different regression algorithms and different feature sets are tried to build a regression model for each emotion dimension(P, A and D), and the performances are measured in terms of correlation coefficient (CF) and R^2 statistics both developed by Karl Pearson. They are defined as follows:

$$CF_{XY} = \frac{\sum_{i=1}^{N} (R(X_i) - \overline{R(X_i)})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{N} (R(X_i) - \overline{R(X_i)})^2} \sqrt{\sum_{i=1}^{N} (Y_i - \overline{Y})^2}} \tag{4}$$

$$R^2{}_{XY} = 1 - \frac{\sum_{i=1}^{N} (Y_i - R(X_i))^2}{\sum_{i=1}^{N} (Y_i - \overline{Y})^2} \tag{5}$$

Where Y_i is the emotion label, $R(X_i)$ is the regression value of feature vector X_i.

To ensure the validity of the results, we use 5-fold cross validation to evaluate the performance of regression models. The dataset is randomly broken into five subsets of the same size, with four being used for training and one for testing, and this process is repeated 5 times and finally the mean CF and R^2 value is taken.

4.2 Comparison of AdaBoost.RM with Baseline Algorithms

Because of the different used datasets, it is not reasonable to compare our approach with existing ones. However we compare our approach with three baseline algorithms. The first is LinearRegression [18] which uses linear regression for prediction, the second is SMOreg [19] which implements support vector machine for regression, and the last is original AdaBoost.R2 algorithm. The experiments are based on the three-modality feature set(A+M+L) introduced in Section 2.3, and the results are showed in Table 4.

Table 4. Performance of our approach compared with that of baseline ones.

	P		A		D	
	CF	R^2	**CF**	R^2	**CF**	R^2
LinearRegression	0.688	0.476	0.823	0.693	0.715	0.529
SMOreg	0.692	0.48	0.828	0.696	0.72	0.542
AdaBoost.R2	0.536	0.284	0.778	0.61	0.672	0.435
AdaBoost.RM	**0.702**	**0.488**	**0.843**	**0.708**	**0.755**	**0.558**

Table 4 shows that among all the regression algorithms our approach achieve the best performance for the regression of all the emotion dimensions(P, A and D), this indicates the effectiveness of our approach. It's to be noted that our approach performs better than AdaBoost.R2, which demonstrates the effectiveness of our modification to AdaBoost.R2. On the other hand we can see that all the regression algorithms have achieved promising performance on our three-modality feature set, which indicates that our feature processing technique and the selected features are really effective to music emotion regression.

4.3 Contributions of Different Modality and Effectiveness of Multi-modal Feature Combination

We conduct a series of experiments to explore the contribution of different modality to the regression of each emotion dimension, and demonstrate the effectiveness of multi-modal feature combination.

The seven feature sets introduced in Section 2.3 are employed for the regression of all the emotion dimensions(P, A and D). The results are showed in Table 5.

Table 5. Contributions of individual modality and effectiveness of multi-modal features.

		P		A		D	
#	**Feature Set**	**CF**	R^2	**CF**	R^2	**CF**	R^2
1	**A**	0.473	0.166	0.724	0.516	0.667	0.38
2	**M**	0.541	0.305	0.823	0.685	0.73	0.508
3	**L**	0.623	0.383	0.461	0.202	0.575	0.328
4	**A+M**	0.571	0.285	0.812	0.663	0.719	0.474
5	**A+L**	0.681	0.465	0.745	0.554	0.728	0.53
6	**M+L**	0.68	0.469	0.828	0.681	0.738	0.542
7	**A+M+L**	**0.702**	**0.488**	**0.843**	**0.708**	**0.755**	**0.558**

In Table 5, the 1st to 3rd rows show that:

1. Among the three modalities, lyric has the biggest contribution to the regression of emotion dimension P.
2. To the regression of dimension A and D, audio and MIDI contribute more than lyric, and MIDI has the biggest contribution among the three modalities. This indicates that MIDI features contain more useful information related to emotion dimension A and D compared with audio and lyric features.
3. Audio and lyric are complementary on the regression of dimension P and A, the reasons maybe that audio signal contains a large amount of energy related information such that the extracted audio features reflect emotional intensity more directly, while lyric contains more semantic information so as to express emotion more directly.

The 4th to 7th rows show that:

1. The regression performance has been enhanced on all the emotion dimensions when any two modalities are combined.
2. The best regression performance has been achieved on all the emotion dimensions when all the three modalities are combined. This indicates that the three modalities provide useful and complementary information for music emotion regression, and the greatest improvement of performance can be achieved when all the three modalities are used.

Generally Speaking, audio signal contains a large number of energy relevant information which reflects emotional intensity more directly; MIDI data contains more information which reflects the concept of music more directly; lyric includes more semantic information which describes emotional inclinations more directly. Audio and MIDI have big contribution to emotion dimension A and D, while lyric has big contribution to emotion dimension P. The three modalities provide complementary information for music emotion regression, and the greatest improvement of regression performance can be achieved when all the three modalities are combined.

5 Conclusion and Future Work

In this paper, we present three main parts of our research work on music emotion regression. First we demonstrate the effectiveness of our regression approach, and then we expound the contribution of each modality to the regression of each dimension of PAD model, and last we verify the performance improvement of emotion regression models brought about by the combination of multi-modal features.

There are two focuses in the future, one is to find more informative features for music emotion recognition, and the other is to build a music emotion retrieval system based on our regression model, in which songs can be retrieved by specifying an emotional state.

References

1. Y. Feng, Y. Zhuang, and Y. Pan: Popular Music Retrieval by Detecting Mood. In: Proc. ACM SIGIR, pp. 375--376(2003)
2. Russell and James A: A Circumflex Model of Affect. Journal of Personality and Social Psychology, vol.39, no.6, pp.1161--1178(1980)
3. Yi-Hsuan Yang, Yu-Ching Lin et al: A Regression Approach to Music Emotion Recognition. IEEE Transactions on Audio, Speech, and Language Processing, vol.16, no. 2(2008)
4. Mehrabian, A.: Framework for A Comprehensive Description and Measurement of Motional States. Genetic, Social, and General Psychology Monographs, vol. 121, pp. 33--361(1995)
5. C.Laurier and J.Grivolla and P.Herrera: Multimodal Music Mood Classification using Audio and Lyrics. In: Proceedings of the 7th International Conference on Machine Learning and Applications(2008)
6. Xiao Hu.et al: Lyric Text Mining in Music Mood Classification. In: ISMIR 2009 Conference Proceedings, pp.411--416(2009)
7. Tzanetakis G., Ermolinskyi, A., and Cook, P: Pitch Histograms in Audio and Symbolic Music Information Retrieval. In: ISMIR 2002 Conference Proceedings, pp.31--38, Paris: IRCAM(2002)
8. Yi-Hsuan Yang et al: Toward Multi-modal Music Emotion Classification. In: Proc. PCM, pp. 70--79(2008)
9. Xiao Hu.et al: Improving Mood Classification in Music Digital Libraries by Combining Lyrics and Audio. In: Proc. of the 10th Annual Joint Conference on Digital Libraries, New York, USA(2010)
10. Q.Lu et al: Boosting for Multi-modal Music Emotion Classification. In: ISMIR 2010 Conference Proceedings, pp.105--110(2010)
11. Lang P. J.: Behavioral Treatment and Bio-behavioral Assessment: Computer Applications. In: J.Sidowski, J. Johnson, & T. Williams (Eds.), Technology in Mental Health Care Delivery Systems. pp.119--137. Norwood, NJ: Ablex(1980)
12. McEnnis, D., C. McKay, and I. Fujinaga: jAudio: A Feature Extraction Library. In: Proc. of the International Conference on Music Information Retrieval(2005)
13. McKay, C., and I. Fujinaga: jSymbolic: A Feature Extractor for MIDI Files. In: Proc. of the International Computer Music Conference(2006)
14. M. A. Hall: Correlation-based Feature Subset Selection for Machine Learning. Hamilton, New Zealand(1998)
15. Weka: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/.
16. Drucker H.: Improving Regressors using Boosting Techniques. In: Proceedings of the Fourteenth International Conference on Machine Learning, pp. 107--115, Morgan Kaufmann, Burlington, Mass(1997)
17. Rumelhart.D et al: Learning Internal Representations by Error Propagation. Parallel Distributed Processing Explorations in the Microstructure of Cognition, vol.1, MIT(1986)
18. Douglas C. Montgomery et al: Introduction to Linear Regression Analysis. 4th Edition, Wiley(2008)
19. S.K. Shevade, S.S. Keerthi, C. Bhattacharyya, K.R.K. Murthy: Improvements to the SMO Algorithm for SVM Regression. IEEE Transactions on Neural Networks(1999)

Application of Free Choice Profiling for the Evaluation of Emotions Elicited by Music

Judith Liebetrau[1,2], Sebastian Schneider[1] and Roman Jezierski[1]

[1] Ilmenau University of Technology, Ilmenau, Germany
[2] Fraunhofer IDMT, Ilmenau, Germany
Judith.Liebetrau@tu-ilmenau.de

Abstract Music evokes and carries emotions. Despite many studies having investigated the relation between music and emotion, current research lacks a systematic and empirically derived taxonomy of musically induced emotions [1]. This work contributes to the question which musical features in particular are able to induce emotions while listening. Problems of defining and measuring emotions are explained. A method to measure affective states induced by music with the help of Free Choice Profiling (FCP) is outlined. Two FCP experiments, assessing the usefulness of the method for emotional research and the selection of test stimuli are described. The shown results are in line with psychological theories of emotions, i.e., the valence/arousal model.

Keywords: Free Choice Profiling, Measuring Emotions, Self-report.

1 Introduction

A diverse range of studies was carried out in the past to investigate how and in which way music influences the emotions of the listener, but still two main questions remain: What exactly is an emotion and how can it be measured? This paper contributes to the question which musical features in particular are able to induce emotions while listening; the research was conducted within a project funded by the German Research Foundation.

A broad overview of common measurement methods can be found in [2]. Although the term emotion is frequently used in literature, authors disagree on its definition, and a simple definition cannot be given. Scherer [3], for example, defines an emotion as an affective phenomenon, distinguishable from feelings, moods, or attitudes. Emotions, resulting from cognitive processes, are necessary for comprehension and appraisal of stimuli on the basis of knowledge. Seeing emotion as a phenomenon consisting of five components (cognitive, neurophysiological, motivational, motor expression, and subjective feeling), Scherer concludes that a universal measure would only become possible by taking into account changes of all components.

Due to the lack of an all-embracing measurement method, each component is measured on its own. The subjective experience of emotions can be assessed in different ways. One possibility is the measurement of changes in psychophysiological parameters like heart rate, heart rate variability and skin conductance during music perception. Numerous studies about the measurement of such physiological correlates

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

of emotions were carried out over the years (a brief review of these methods can be found in [2,4,17]), but this paper is focused on a second possibility: The assessment of the subjective experience of emotions during music perception based on self-reports [8]. In self-report methods, the subjects are stimulated to verbalize and express their emotions towards stimuli. Different techniques, called answering formats, exist to assess the participants' emotions, such as affective scales, free descriptions, or the use of "emotional space" [4]. Every answering format has different advantages and drawbacks when measuring emotions. The next section will explain in more detail advantages and disadvantages of common measuring methods.

2 Challenges in Measuring Subjective Experience of Emotions via Self-reports

In [2, p. 210] Zentner states that "there are four important limitations to self-report methodology […]: a) demand characteristics, b) self-presentation biases, c) limited awareness of one's emotions, and d) difficulties in the verbalization of emotion perception […]".

While the assessment of subjective experience with closed-response self-report methods such as adjective scales [5] or emotional spaces [6] is ensuring efficiency and, to some degree, a standardization of data collection, "the predetermined choices [of descriptors] might influence the participant to respond along the provided categories [and] the interpretation of the terms provided by the researcher might vary considerably across people […]" [2, p. 193]. One attempt to overcome the problems of closed-responses is the usage of a free response measurement: Subjects are allowed to explain the nature of the state they experience, i.e., an emotion while listening music, in their own words, for example in written form or an interview. A content analysis of the narrative establishes the link between music and the induced emotions. Unfortunately, the data treatment and interpretation of such a content analysis is not an easy task and cannot be automated. A second disadvantage lies in the different linguistic abilities of the subjects – some might lack an appropriate vocabulary to describe the emotional experience during listening to music. This might lead to a loss of information. A possible way to promote the advantages of both measurement approaches is the combination of open and closed-response format.

An approach using an open response format in combination with a closed-response format was presented in [8], the so-called Free-Choice Profiling (FCP). By applying FCP, subjects first define and identify individual attributes (emotional terms, also called descriptors) by themselves. The rating of intensity of the emotional experience during music perception is then done with the help of adjective scales, where the individual attributes are used as labels. Due to the design of the test method, it is taken into consideration that different subjects might use terms in different ways, or different terms with the same underlying meaning. The study mentioned in [8] was able to obtain clear and interpretable results consistent with music theory and emotional psychology. However, the study investigated only a small set of major/minor chord items, and as it was the first application of FCP in the field of emotions in music, questions of reliability and general feasibility of the method remained.

3 Experimental Design and Parameters

The promising results of [8] led to a research project funded by the German Research Foundation to verify FCP as a useful test methodology for the assessment of emotions and to enable a better classification of musical parts based of their emotional impact on music perception.

This paper presents two preliminary FCP studies of this ongoing project that were conducted with different scopes: The first experiment aimed at assessing the selection of suitable test stimuli, in terms of their degree of emotional impact. The target of the second study was to verify the usage of FCP as a suitable test methodology by using different test material. The test method, items, and participants for both experiments are explained in this section.

3.1 Test Method in General

FCP, a method common in food research, was used to identify individual attributes (emotional terms) and to rate the liking and/or intensity of those attributes. The procedure, which is outlined in detail in [8, 9], helps to identify significant attributes, discrimination, and panelist performance. It takes into consideration that different subjects might use terms in different ways, or different terms with the same underlying meaning. In recent years, FCP was also successfully applied and refined in the field of user experience to assess multimodal quality perception [18].

As mentioned in [18] the FCP is structured into four different parts, referred to as *introduction*, *attribute elicitation*, *refinement of attributes,* and *sensory evaluation*. In the *introduction* the nature of descriptive evaluation, in particular the use of the participant's own attributes to describe the perceived emotionality of test items, is explained in detail. This first step of the method is the most crucial part, because here the cornerstone for the assessment is laid. Subjects have to understand the method correctly, but special care must be taken not to influence them in a certain direction. Therefore, the participants are shown how to find attributes that define emotions with an easy task of a different perceptual domain[1]. The *attribute elicitation* aims at finding individual emotion attributes that characterize each participant's emotional perception of the different test stimuli. In this study, participants listened to a small representative subset of test items (see Section 4.2) and wrote down the perceived emotions using their own words, without any limitation concerning the number of attributes. No additional technique like repertory grid method [10] or natural grouping [10] was used as support for the elicitation of attributes. In the third step a *refinement of attributes* was done. Here, strong attributes were chosen out of all developed attributes according to two rules: First, attributes must be unique and each attribute must describe only one aspect of emotion. Second, the participants must be able to define the attribute in their own words. Hence, the participants had to write down a definition of each of the attributes left over for the final evaluation. For the *sensory evaluation* all generated attributes were printed out on paper together with 10 cm long

[1] For example, as it was the case in this study, the participants are asked to describe the emotional impact of different movies or photos.

scales, labeled with “min” and “max”[2]. These individual score cards, one for every test item, were used for the evaluation of all test items, which were presented randomly one by one. The subjects were advised to mark the perceived strength of each attribute for each test item.

3.2 Test Items

Experiment 1: The test items consisted of eight specifically designed major/major and minor/minor chord combinations, derived from the circle of fifths (see Figure 1). Each item consisted of two chords played one after the other: C/F, C/G, C/B, C/D♭ (major), as well as c/f, c/g, c/b, c/d♭ (minor). These are the two chords located next to C (F and G), and the ones furthest away (B and D♭). To assess the selection of suitable test stimuli, in terms of their degree of emotional impact, these musical phrases were also varied in instrument choice and tempo. Their length varied, depending on instrument choice and tempo, from approx. 2.5s to 7.5s. The decision to use these basic musical structures was made in order to exclude as many other variables as possible, including familiarity with well-known musical pieces.

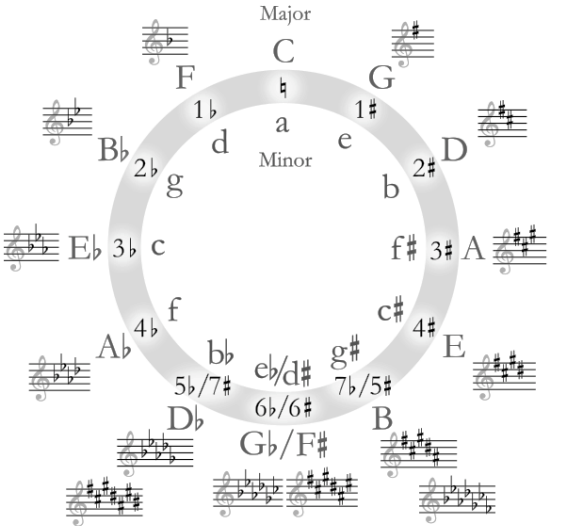


Fig. 1. Circle of fifths. (from: en.wikipedia.org, licensed under CC BY-SA 3.0)

Three different instruments were used: Violin, piano, and synthesizer. Violin was chosen because of the possibility to induce sad emotions [12]. Former studies indicate that artificial instruments, e.g., a synthesizer, can lead to a decreased recognition of sad emotions [12]. Piano was chosen because of its broad usage and popularity in other studies, i.e., to allow comparability [13]. Furthermore, three different tempi (30,

[2] Where “min” means that the attribute is not perceived at all, while “max” refers to its maximum pronounced form.

70, and 120 bpm) were used as conditions. The previous FCP study [8] used 30 bpm only and indicated that this tempo already induces a slightly sad emotional offset, so this tempo was used in this study for comparison. 70 bpm was chosen as it is close to resting heart rate and can be seen as a "normal" state of activation for the listener. 120 bpm is a common tempo for modern dance music as well as for modern marches ("march tempo") and can be regarded as more activating. This results in a total set of 72 musical phrases.

Experiment 2: Several participants of experiment 1 mentioned that the test items appeared to be too short for eliciting distinctive emotions. The second experiment therefore aimed to investigate whether longer test items were perceived differently. New items were created according to Table 1. As the first experiment took the participants nearly 60 min. on average, it was decided to use fewer items to compensate for the longer item length. Only two different tempi (70 and 120 bpm) and only one instrument (synthesizer) were used. The item length ranged from 6s to 10s. The reduced set of 16 test items led to a much shorter rating time of only 15 min. on average.

Table 1. Chord combinations used in the second experiment.

Item number	Chord combination
1	C-G-C-F-C-G-C-F-C
2	a-e-a-d-a-e-a-d-a
3	C-A-C-D-C-A-C-E-C
4	a-f#-a-b-a- f#-a-c#-a
5	C-F#-C-B-C-F#-C-D♭-C
6	a-d#-a-g#-a-d#-a-b♭-a
7	C-E♭-C-B♭-C-E♭-C-A♭-C
8	a-c-a-g-a-c-a-f-a

3.3 Test Panel

In the first test 24 subjects, 9 female and 15 male, participated. The average age was 24.8 years.

The number of subjects in the second experiment was 10, with an average age of 24.7 years. Half of the participants were male and the other half female.

Any participant took part in only one of the experiments. Although some subjects reported slight hearing damages, none of the subjects were rejected from test participation and analysis.

4 Experiment 1

The first experiment aimed at assessing the selection of suitable test stimuli and a general re-evaluation of FCP for the assessment of emotional impact while listening to musical phrases.

4.1 Test Facilities

The tests were conducted in the Audio Lab at Ilmenau University of Technology, a room compliant to ITU-R BS.1116-1 [14], to EBU 3276, and to DIN 15996. Its exact dimensions are 8.4m x 7.6m x 2.8m. Two identical Genelec 1030A loudspeakers were used in the test, placed on stands at ear height of the seated subjects. Participants were seated in the sweet spot position in front of a desk with a flat screen monitor, keyboard, and mouse. The arrangement of the speakers and the listening positions are in accordance with ITU-R BS.1116-1.

4.2 Test Procedure in Detail

Introduction: Each participant received a short introduction about the test in general and the test method FCP. They were handed out a privacy policy and had to fill out a short questionnaire regarding demographics, musical knowledge, and their current mood. For a better understanding of the attribute elicitation and listening task, each subject was asked to imagine two different (known) movies and to verbalize the differences in the emotions they associated with them. The supervisor took care to avoid giving predetermined attributes that might influence them in a certain direction.

Attribute Elicitation: During this stage, each participant assessed a representative selection of 16 of the final test items, that is one item for each instrument, tempo, and key, and wrote down the verbal descriptors with which they would have to rate these items in the fourth part (attribute rating). A graphical user interface (GUI) was used, allowing the subjects to listen to each item as often as they wanted. During this part, the supervisor left the room for the control room, in order not to disturb the participant. The participants were seated in a 90° position to the control room window, thus the supervisor remained available, either via eye contact or a microphone connection.

Attribute Refinement: After the participant signaled that he/she was done, the supervisor and the participant reviewed the attribute list together. The participant decided if some words could be summarized to one single term or should be renamed. After this, the participant was asked to give a brief explanation for each term, if possible. The attribute list was reviewed once again afterwards.

Attribute Rating: Starting with a short rating test of 3 items and all of his or her attributes, each participant carried out a training task. In case the participant felt the need to apply changes, they were allowed to modify their descriptors one last time.

After this, the actual test started, where each subject rated all 72 items with the complete set of their descriptors. The test allowed the participants to listen to each test item as often as they wanted, but it was not allowed to revise ratings of prior items. It was planned to have a rating software right from the beginning, but due to a computer failure the first 4 subjects did a rating on paper with a list of their attributes on the left and for each attribute a 10 cm long rating scale on the right side of the sheet. The scales were labeled with min and max. Later subjects carried out the rating with software. The design of the graphical user interface was similar to the rating sheets. If

participants took longer than 60 minutes, they were asked to take a break of approximately 10 minutes before they went on.

4.3 Results

The data was analyzed with a Multiple Factor Analysis (MFA) [15], a widely used method in sensory profiling. As each participant uses his/her own vocabulary, a multi-dimensional perceptual space – the verbal descriptors representing the dimensions – is created. MFA is very similar to Principal Component Analysis (PCA)[3]: it compares the individuals' perceptual spaces and combines them into a single global one. An MFA provides mainly two outputs: a) The mean location of the test items on the global space and b) the location of the verbal descriptors on these dimensions.

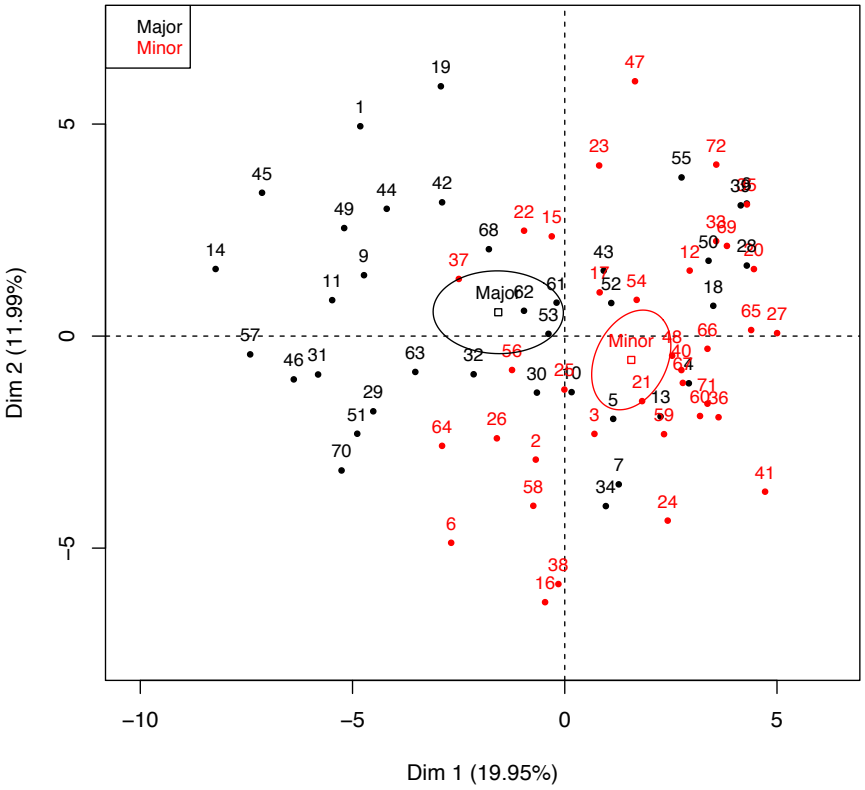


Fig. 2. Graph of the first two dimensions of experiment 1 with an explained total variance of 32%. Shown are the major and minor chord-combination groups and their respective confidence ellipses for the mean of each group.

[3] In fact an MFA computes nested PCAs.

Figure 2 shows a graph of the first two dimensions with the mean location of the test items and confidence ellipses for the means of major and minor categories. The non-overlapping ellipses clearly indicate that these categories were rated significantly different. In total, the first two dimensions declare only 32% of the total variance of the original data. One reason for this could be that there was little agreement among participants.

Still the arrangement of the test items on the first two dimensions is sensibly interpretable in several other ways beyond key[4]: All chord-combinations of each key, except the ones featuring B and D♭, were rated significantly distinguishable and were ordered from left to right on dimension 1 according to their distance to C on the circle of fifths (see Figure 1).

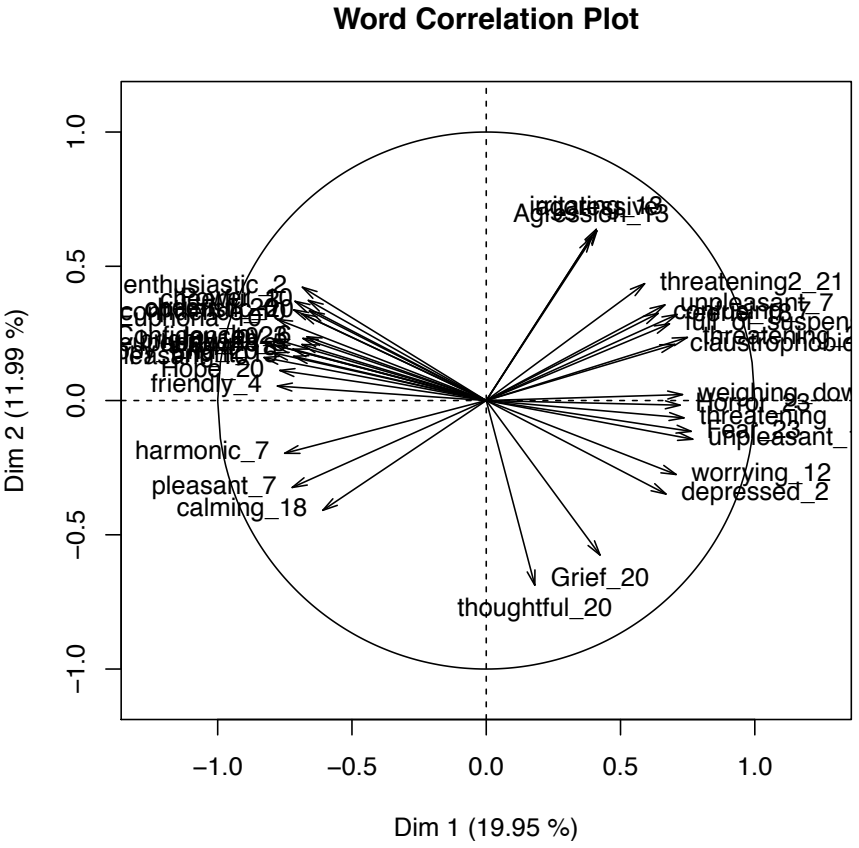


Fig. 3. Word chart of the significant listener descriptors for the first two dimensions of experiment 1. The numbers behind the descriptors refer to the listener.

[4] The respective graphs cannot be shown here due to space reasons, but are available on request.

The second dimension (y-axis) features the faster tempi and the synthesizer-sound on the upper part, while the lower part primarily contains the 30 bpm tempo and violin- and piano-sounds. The instrument synth was rated significantly higher than violin and piano, and 30bpm and 120bpm can be clearly separated in the second dimension.

To identify the perceived emotions the participants associate with these dimensions, Figure 3 shows the respective word chart of the first two dimensions. Only those descriptors contributing to both dimensions with an $R^2 \geq 0.5$ are plotted, hence not all descriptors of all participants are present. All verbal descriptors were originally given in German and translated by the authors, the English translations shown here may hence convey slightly different meanings. Word charts tend to be crowded, therefore Table 2 gives an overview of these attributes, ordered by participant.

The first dimension (x-axis) features positive descriptors on the left side (excited, happiness, confidence, euphoria, harmonic, pleasant, calming, etc.), and negative ones on the right side (fear, horror, menacing, aggressive, irritating, unpleasant, depressed, etc.). This conforms very well with the concept of "valence" in emotional psychology [2, 4]. The second dimension (y-axis) does not contain many descriptors (which results in its low explained variance), but they are clearly interpretable: the lower part shows descriptors of low activity, such as: calming, pleasant, harmonic, but also grief, thoughtful, depressed and unpleasant. The upper part contains descriptors that are clearly active, for example, aggressive and excited. This again conforms with another well-known concept: arousal [2, 4].

Table 2. Significant descriptors contributing to dimensions 1 and 2 of experiment 1. The numbers behind the descriptors refer to the listener.

Attribute	Attribute	Attribute
threatening	claustrophobic_7	calming_18
aggressive	Euphoria_10	light_18
depressed_2	Joy_10	cheerful_20
cheerful_2	Power_10	Hope_20
enthusiastic_2	self-confidence_10	thoughtful_20
friendly_4	pleasant_10	Happy_End_20
full_of_suspense_4	unpleasant_10	Grief_20
weighing_down_4	bright_12	optimistic_20
euphoric_5	worrying_12	threatening_21
Joy_6	Agression_13	threatening_21
pleasant_7	irritating_13	Confidence_23
harmonic_7	bright_15	Fear_23
confusing_7	euphoric_18	Horror_23
unpleasant_7	eerie_18	

5 Experiment 2

Several participants of experiment 1 mentioned that the test items appeared to be too short for eliciting distinctive emotions. To assess the effect of longer test items a second experiment was conducted. The two experiments are comparable in their procedure, but in this second experiment we made a slight change to the preparation task for the *attribute elicitation*. The attribute election procedure itself stayed the same. Minor changes were the use of a different test facility room and test items.

5.1 Test Facilities

The tests were conducted in the Audio Lab at Fraunhofer IDMT compliant to ITU-R BS.1116-1 [14], to EBU 3276 and to DIN 15996. Its exact dimensions are 6.90 x 4.60 x 2.70. Two identical K&H O-510 loudspeakers were used in the test, placed at ear height of the seated subjects. Participants were seated in the sweet spot position. The arrangement of the speakers and the listening positions are in compliance with ITU-R BS.1116-1. The changes in test facilities are considered not to bias the results. The room characteristic is in line with the room characteristics of the first experiment. Although the test equipment is not exactly the same like experiment 1, the same class of high quality loudspeaker was used for the tests.

5.2 Test Procedure in Detail

The general procedure of this experiment was very similar to the first experiment (see section 4): The introductory task of imagining two movies was replaced, because for some participants the task was too abstract and they had problems understanding the intention of the attribute elicitation task. Instead, participants were now handed out five different images[5], which were taken from the International Affective Picture System (IAPS)[6] database. They were asked to explain what emotions these images elicited and to verbalize the similarities and dissimilarities.

5.3 Results

Experiment 2 was analyzed in the same manner as experiment 1 (cf. Section 4.3). Figure 4 shows the graph of the mean location of the test items on the first two dimensions. It is apparent that the explained variance is higher (42.6%) than in experiment 1 (32%), which can be interpreted as a slightly higher agreement among the participants on what they perceive.

[5] The images portrayed: 1) several woodlice, 2) a woman and a child close together, 3) a wolf, 4) a rabbit, and 5) a lonely road through grass-covered plains.

[6] http://csea.phhp.ufl.edu/Media.html

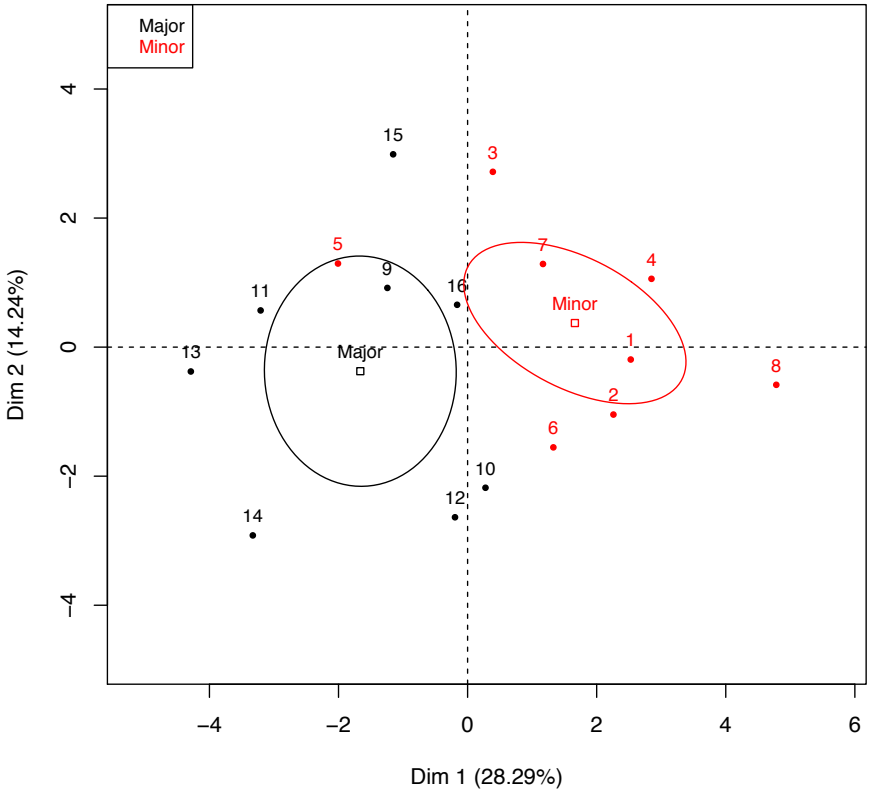


Fig 4. Graph of the first two dimensions of experiment 2 with an explained total variance of 42,6%. Shown are the major and minor chord-combination groups and their respective confidence ellipses for the mean of each group.

Considering the position of the test items on these dimensions, the picture is partially similar to the one of experiment 1: On the first dimension (the x-axis), the left hand side contains all major chords except one, while the right hand side contains all minor chords except one. Furthermore[7], the items are – as it was the case in experiment 1 – sorted according to the circle of fifths, with the neighboring chord-structures on the left hand side and the opposing chord-structures on the right hand side of each key group. The location of the items on the second dimension (y-axis) is not as obvious as in experiment 1, but it can be noted that the items rated most positive on this dimension are the faster ones (120 bpm), while the most negative ones are the slower ones (70 bpm), and that these groups are significantly different. The word chart (Fig. 5, see p. 13) of the first two dimensions matches the item location chart: On dimension 1 (x-axis), positive descriptors can be found on the left hand side: happy, cheerful, euphoria, impressive, heroic, festive, etc.; negative descriptors

[7] As before, the respective graphs cannot be shown here due to space reasons but are available on request.

are located on the right side of the axis: menacing, danger, loneliness, exhausted, thoughtful, etc. Compared to Figure 3 the second dimension (y-axis) is not that clearly marked as in experiment 1, but in general the more "active" descriptors (happy, cheerful, menacing) are located on the positive side of this dimension, while the negative side contains mostly "inactive" descriptors: heroic, festive[8], loneliness, exhausted, thoughtful, etc. Again, Table 3 shows all the descriptors significantly contributing to the dimensions 1 and 2.

Table 3. Significant descriptors contributing to dimensions 1 and 2 of experiment 2. The numbers behind the descriptors refer to the listener.

Attribute	Attribute	Attribute
boring_1	Suspense_4	cheerful_8
not_harmonic_1	monotone_5	Depression_9
heroic_2	menacing_6	happy_9
euphoric_2	ominous_6	exhausted_9
menacing_2	promising_6	Loneliness_10
depressing_2	carefree_6	Euphoria_10
Success_3	delighted_7	festive_10
Danger_3	thoughtful_7	Party_mood_10
annoying_4	sad_7	impressive_10

In summary, the location of the items on these dimensions and the respective descriptors concur with experiment 1 in that the first dimension can easily be interpreted as "valence". In the case of the second dimension, it seems that the participants knew what they wanted to rate, but then had problems to actually discern the items. This is not surprising as the difference in activation between 120 and 70 bpm is clearly much lower than the difference between 120 and 30 bpm, as it was the case in experiment 1. Nonetheless, the second dimension can easily be interpreted as "arousal".

[8] In the case of "heroic" and "festive" it might be argued that these are active descriptors, but the German connotation of the original descriptors is more that of a ceremonial atmosphere.

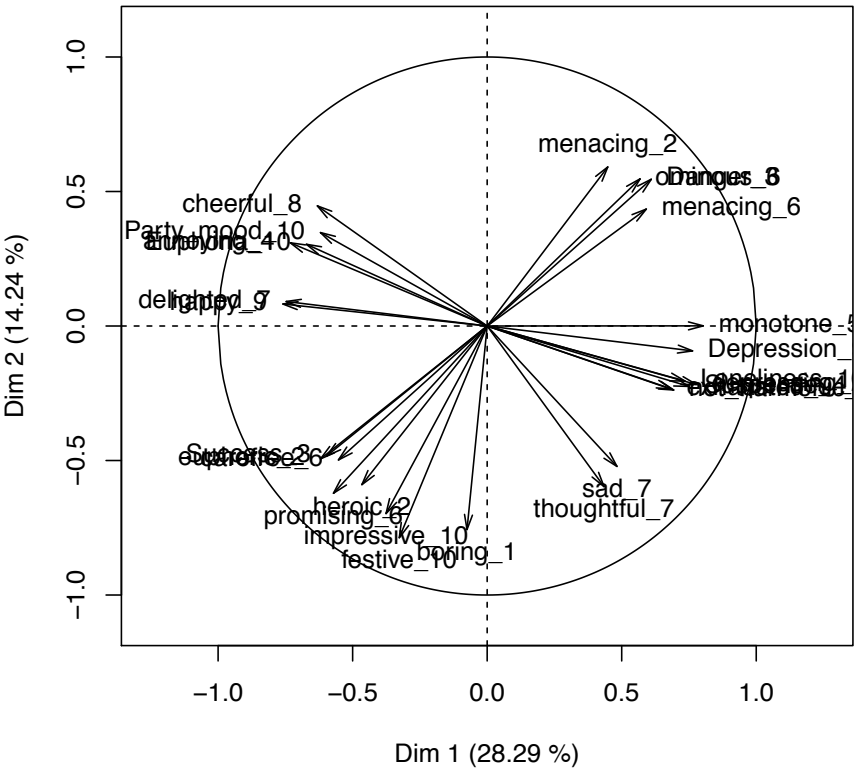


Fig. 5. Word chart of the significant listener descriptors for the first two dimensions of experiment 2. The numbers behind the descriptors refer to the listener.

6 Conclusion and Further Work

In this paper, we propose and investigate FCP as a test method to overcome drawbacks of common self-report methods, to assess the emotional state of a subject during music perception. By applying FCP, subjects define individual attributes (emotional terms) by themselves. The rating of the intensity of the emotional experience during music perception is done with the help of adjective scales, where for each subject their individual defined attributes are used as labels. To prove the feasibility of FCP for the evaluation of emotions elicited by music and to assess the selection of suitable test stimuli, two experiments were carried out.

The results of experiment 1 showed that the subjects rate the emotional impression according to dimensions of valence and arousal, which are commonly proposed by emotional psychology. Furthermore, simple major and minor chord combinations could directly be linked to the dimension of valence, with the participants being able to sort the chord-samples according to the circle of fifths. The second dimension features the faster tempi and the synthesizer-sound on one side, while the other side primarily shows the 30 bpm violin- and piano-sounds. This leads to the conclusion

that the second dimension represents “arousal”. Although the detailed analysis shows clearly interpretable results, the results only declare 32% of the total variance of the system. This further leads to the conclusion that there is rather large disagreement between the participants on what they perceive, and the "least common denominator" is fairly small.

Verbal comments of the participants led to the assumption that the musical phrases were too short to elicit emotion. Experiments examining the lower bound of length in which emotions can be perceived have been conducted, e.g., [19, 20], finding that excerpts as short as 250-500ms are sufficient to elicit emotions. However, these studies examined emotions on a very basic level, reducing the spectrum to a binary happy/sad decision [20] or neutral/moving [19]. Thus, it remains unclear whether participants are able to precisely classify their perceived emotions in a multi-dimensional space with such short pieces. Furthermore, the studies used excerpts of classical and well-known musical pieces. This poses the question whether participants rated their actual perceived emotions or rather their remembered emotions based on familiarity.

To prove the hypothesis that longer stimuli are more easily classified, a second experiment was conducted where longer test stimuli were used. While the items of experiment 1 consisted of two chords with a maximum item length of 7s were used, experiment 2 had items with nine chords per item and a maximum length of 10s.

The explained variance of the first two dimensions in experiment 2 is slightly higher than in experiment 1 with 42.6%, which can be interpreted as a slightly higher agreement among the participants on what they perceive. In general, the results of the first experiment were confirmed. Unfortunately, the extension of the musical phrases did not lead to a significant higher explained variance.

Although the results are easily interpretable and sensible, the low explained variances of the results are puzzling. One explanation could be that emotion is a very subjective experience, which is not easy to describe or indicate.

Furthermore, the test method cannot solve the problem of awareness of an emotion as mentioned in [2, p. 210 et seq.]. When defining an emotion as consisting of several components, it is questionable which part of an emotion is accessible at all and which part is accessed in a self-report. This could be another reason for the low explained variance of the test results.

FCP is able to approach two of the four problems raised by Zentner ([2, p. 210 et seq.], also see Section 2): Because no fixed responses are given, participants do not feel the need to comply with certain emotional concepts and will not feel *demand characteristics*. Secondly, the *difficulties in verbalization of musical emotions* are partly compensated by FCP’s ability to directly compare and group correlating descriptors of all participants. Hence, it is not so important that the participant is able to express emotions with a complex vocabulary, but rather that he/she is able to discern and rate the perceived emotions.

To further investigate the dependency of the linguistic abilities on the rating and see if FCP really solves the addressed problem, we plan to conduct new experiments, using the same test items as in experiment 1. The next experiment will use a pictorial rating system called SAM (Self-Assessment Manikin, Fig. 5) [16], a common and well-researched rating system in emotional research. This rating system assesses the three dimensions valence, arousal and dominance in a non-verbal way and is thus

suited to be used by children and/or non-native speakers. The participants of this experiment will consist of "Amazon Mechanical Turk" (MTurk) workers[9]. This so-called "clickworker"-platform allows to offer easy tasks that can be solved with a few mouse-clicks, such as annotation tasks, to registered workers. Amazon MTurk is a cheap and efficient way to have many test items annotated by a lot of people in order to build a ground truth. The results will be compared to those of the experiments already conducted.

Acknowledgments This work was supported by the German Research Foundation (DFG)-funded project "Fundamental research on the development of algorithms for the analysis of emotion-psychological characteristics of music signals".

References

1. Zentner, Marcel; Grandjean, Didier; Scherer, Klaus R. (2008): "Emotions evoked by the sound of music: Characterization, classification, and measurement. " In: Emotion 8 (4), 494–521
2. Sloboda, John A.; Juslin, Patrik N.; Frijda, Nico H. (Hg.) (2010): Handbook of music and emotion. Oxford: Oxford University Press
3. Scherer, K. R. (2005). "What are emotions? And how can they be measured?" Social Science Information, 44(4), 695–729
4. Nagel, Frederik (2007): Psychoacoustical and psychophysiological correlates of the emotional impact and the perception of music. 1. Aufl. Göttingen: Sierke
5. Hevner, K.: "Experimental studies of the elements of expression in music." American Journal of Psychology, 48, 246–286, 1936
6. Russell, J. Russell: "A circumplex model of affect." Journal of Personality and Social Psychology 39 (1980), pp. 1161–1178, 1980
7. Zentner, Marcel; Eerola, Tuomas (2010): Self-report measures and Models. In: Patrik N. Juslin, John A. Sloboda und Nico H. Frijda (Hg.): Handbook of music and emotion. Oxford: Oxford University Press, 187–221
8. Schneider, S.; Raschke, F.; Gatzsche, G.; Strohmeier, D.: "Free Choice Profiling and Natural Grouping as Methods for the Assessment of Emotions in Musical Audio Signals", 126th AES Convention, 2009 Munich
9. Lawless, H.T., and Heymann, H. "Sensory evaluation of food: principles and practices". Chapman & Hall, New York. 1999
10. Jack, F.R. and Piggott, J.: "Free choice profiling in consumer research," Food quality and Preference, vol. 3, no. 3, pp. 129–134, 1992
11. Williams, A.A., Langron, S.P. "The use of Free-choice Profiling for the Evaluation of Commercial Ports." In: Journal of the Science of Food and Agriculture 35, pp. 558-568, 1984
12. Behrens, Green: "The Ability to Identify Emotional content of Solo Improvisations Performed Vocally and on Three Different Instruments"; Psychology of Music ,Volume 21(1):20-33 (1993)
13. Hailstone, Julia; Omar, Rohani; Henley, Susie; Frost, Chris; Kenward, Michael; Warren, Jason: "It's not what you play, it's how you play it: Timbre affects perception of emotion in

[9] https://www.mturk.com/mturk/welcome

music." In: The Quart. J. of Expt. Psych (The Quarterly Journal of Experimental Psychology), 1962, No. 11, 2141–2155, 2009
14. Recommendation ITU-R BS.1116-1 (10/1997) Methods for the subjective assessment of small impairments in audio systems including Multichannel Sound Systems. International Telecommunication Union, Radiocommunication Assembly
15. Abdi H. and Valentin, D.: "Multiple factor analysis (mfa)," Encyclopedia of measurement and statistics, pp. 657–663, 2007.16
16. Bradley, Margaret M.; Lang, Peter J. (1994): Measuring Emotion: The Self-Assessment Manikin and the Semantic Differential. In: Journal of Behavior Therapy and Experimental Psychiatry (Vol. 25, No. I.), S. 49–59
17. Stemmler, G. (2003): ''Methodological Considerations in the Psychophysiological Study of Emotion'', In: R.J. Davidson, K.R. Scherer and H. Goldsmith (eds) Handbook of the Affective Sciences, pp. 225–55. New York and Oxford: Oxford University Press.
18. D. Strohmeier, S. Jumisko-Pyykkö, and K. Kunze, "Open profiling of quality: a mixed method approach to understanding multimodal quality perception," Advances in Multimedia, vol. 2010, Article ID 658980, 28 pages, 2010.
19. S. Filipic, B. Tillmann, and E. Bigand, "Judging familiarity and emotion from very brief musical excerpts," *Psychonomic Bulletin & Review*, vol. 17, no. 3, pp. 335–341, Jun. 2010.
20. I. Peretz, L. Gagnon, and B. Bouchard, "Music and emotion: perceptual determinants, immediacy, and isolation after brain damage," *Cognition*, vol. 68, no. 2, pp. 111–141, 1998.

SUM: from Image-based Sonification to Computer-aided Composition

Sara Adhitya[1] and Mika Kuuskankare[2]

[1] EHESS/ IUAV/ STMS: IRCAM/CNRS/UPMS
sara.adhitya@ehess.fr
[2] Sibelius Academy / CCRMA Standford University
mkuuskan@siba.fi

Abstract. This paper will discuss the development of the SUM tool, a user library with a graphical user interface within the computer-aided composition environment of PWGL, aimed at the integration of image and sound. We will discuss its internal structure, consisting of image layers, mappers, and paths. We will explain the mapping process, from the retrieval of graphic data to its translation into audio parameters. Finally, we will discuss the possible applications of SUM in both image sonification and computer-aided composition, resulting from this structure.

Keywords: image sonification, graphical computer-aided composition, open graphic score, structure

1 Introducing the SUM Tool

The SUM tool allows the integration of image and sound through a graphic user interface. It was originally developed as an audio-visual representation tool in urban planning [1], an applied discipline involving the spatial composition of temporal systems. The traditional use of multiple 2-dimensional graphic maps makes it difficult to represent dynamic flows, as well as synthesise multiple layers due to legibility constraints. Thus SUM provides a more temporal approach to spatial composition through sonification – the representation of data through auditory means [2].

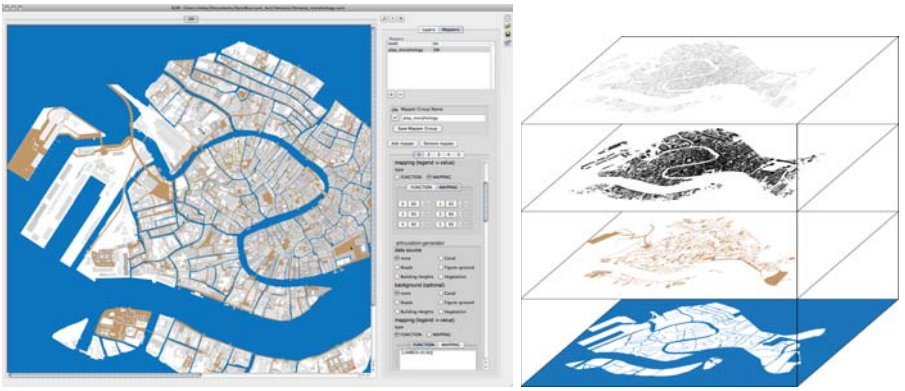


Fig. 1. The sonification of multiple urban maps in the SUM tool

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

Due to its design application, SUM supports both the importation and creation of multiple image layers (raster and vector) as data input. This data is then retrieved through the drawing of one or more vector paths over the areas of interest, and their graphic attributes mapped to sound attributes results in the generation of audio parts. Thus SUM supports a multi-dimensional spatio-temporal approach to image sonification, which sets it apart from other image sonification toolkits such as SonART [3].

As a user library within PWGL [4], a widely-used Lisp-based visual computer-aided composition environment, SUM can also be used as a graphical composition tool. PWGL's internal music notation editor strictly allows the description of object-based graphical scores [5], rather than the pixel-by-pixel exploration of a score as an image. Other graphical computer-aided composition environments, such as HighC [6] and Iannix [7], inspired by Xenakis' UPIC system [8], support the drawing of graphic objects but are limited to a single horizontal time axis. However SUM, with its ability to create and read objects along multiple spatio-temporal paths, allows an image to be composed and played as an open graphic score from multiple perspectives.

This paper will discuss the structure of SUM, which supports a multi-dimensional approach to both image sonification and graphical computer-aided composition.

2 The Structure of SUM

The SUM tool consists of three main components: images; paths; and mappers. The following section will explain each of these components and their inter-relationships.

2.1 Images

SUM uses images as data-sources. Each image is described by a 'color-key', in which each color of interest is allocated an arbitrary numerical value, to be referenced in the sonification mapping process. SUM supports the superimposition of multiple images, which allows the synthesis of overlapping graphic information, visualisable as a '3D' matrix of data as shown in figure 2. A group of data-sources is called a 'dataset', from which any number of image layers may be drawn upon as data-sources in the mapping process.

SUM allows the co-existence of raster and vector images. The flexibility of raster importation permits any visualization, including that produced by other software, to be sonified. The tool's vector drawing ability allows it to be used as a computer-aided design tool, such as Adobe Illustrator or AutoCAD, with graphic changes able to be made internally.

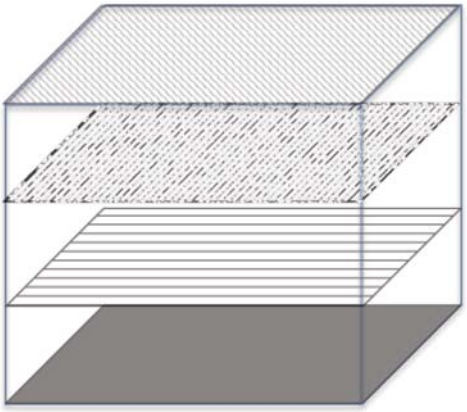


Fig. 2. Visualisation of a 'dataset' of 2D images as a 3D matrix

2.3 Paths

A path is responsible for defining the connection between the graphic space and musical time. It is a spatio-temporal object consisting of the following qualities: location; direction; delay; duration; and speed. The path is drawn as a vector polyline by the user over the area of interest, and then assigned a speed and delay. SUM supports the co-existence of multiple paths of various speeds and delays.

2.2 Mappers

A mapper is responsible for defining the sound output of the mapping process. It translates the graphic attributes retrieved from the image into discrete audio events, defining the sound attributes of pitch, volume, articulation and timbre. The definition of each sound attribute is independent of another. Thus one mapper can refer to multiple data-sources. A group of mappers is termed a 'mapper-group'.

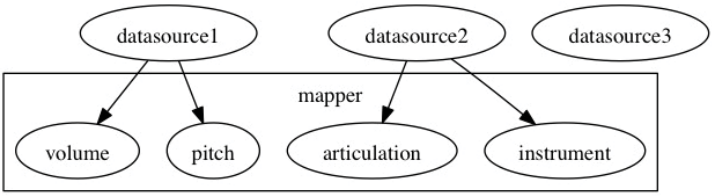


Fig. 2. The SUM mapper: one possible definition of sound attributes by data-source

3 The SUM Mapping Process

The SUM mapping process from image to sound is a two-fold process: graphic data is retrieved from a data-source by a path; it is then applied to a mapper for transformation into audio attributes.

3.1 Data Retrieval

The SUM mapping process is path-driven. Data is retrieved through the drawing of a vector path on an image, and the sampling of the image along this path. The vector path is rasterized according to Bresenham's line algorithm [9] in order to break it down into discrete sampling points, while retaining the order of the points to determine the direction of the path along which the time progresses. Thus for a line extending upwards and to the left, the pixels would be sampled in the order shown in figure 3.

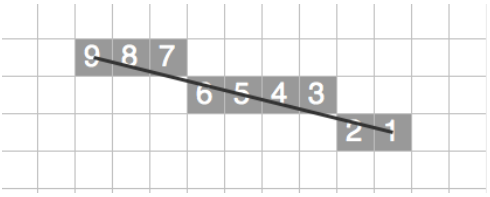


Fig. 3. Diagram of Bresenham's line algorithm, showing sampling order

Each raster map image is then sampled pixel-by-pixel to retrieve the data of interest per each sample-point along the path. The user-defined start-time and playback speed determines the temporal structure of the mapping process.

3.2 Parameter-Mapping

After retrieval of the graphic information along a path, these values can be applied to a mapper in order to generate the desired sound attributes of an acoustic signal (pitch, volume, articulation, and timbre). The parameter-mapping process is defined by assigning a legend, from a given data-source, with a sound value. This can be implemented either directly through the graphic user interface or by using Lisp for more complicated mappings.

Application of a path to a mapper produces a set of sound parameters, which can then be used to drive a wide-variety of internal or external instruments. PWGL has its own internal synthesizer as well as MIDI and OSC output. This allows connection to external sound synthesis engines such as Max/MSP and flexible possibilities for sound output.

It should be noted that a path and a mapper are independent of each other in terms of data-source/s. Thus different mappings can be generated from the same dataset of data-sources.

4 The SUM Compositional Process

This section will relate the SUM process to the compositional process. Here we introduce the concept of the SUM score, consisting of multiple SUM parts.

A SUM part is a sequence of audio events, the qualities of which are defined by the retrieval of data from an image with a path, and applying this path to a mapper. Thus the generation of a SUM part is a path-driven process. Application of multiple paths to one mapper will produce multiple SUM parts of the same timbral quality, but of variable temporal structure. Application of the same path to multiple mappers will produce multiple SUM parts of the same spatio-temporal quality, but of variable timbral qualities. Different combinations of paths and mappers allow the generation of numerous SUM parts from the same dataset. Figure 4 shows one possible network of paths and mappers producing a SUM score.

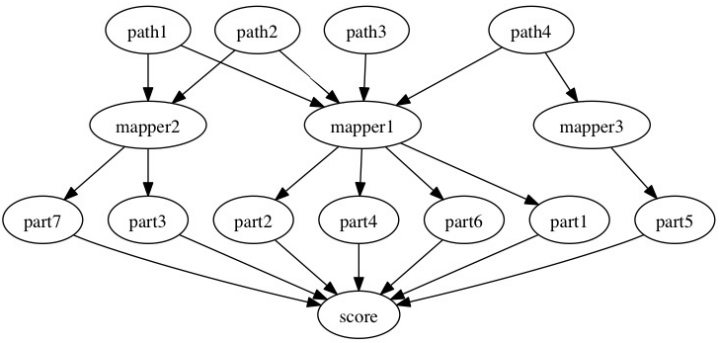


Fig. 4. An example of a SUM score - one possible network of paths and mappers

5. Applications

The flexibility of the mapping process established between image and sound has the potential for application in both image-based sonification and computer-aided composition.

5.1 Image Sonification – Playing of 'Visual Music'

The SUM tool, with its image-based input and user-defined mapping process, supports the sonification of any color-coded image. This means that any bitmap image can be sonified according to its own color-key.

One artistic application is in the playing of 'visual music' – the generation of musical concepts such as rhythm through graphic means. One visual composition technique is through the spatial arrangement of colour, as explored by Piet Mondrian in his series of paintings entitled 'Composition' utilizing the primary colours of red, yellow and blue. Here we demonstrate the sonification of his work *Woogie Broadway*

Boogie (1942-43), in which he attempted to express the musical rhythm of the 'boogie woogie' through colour, and in addition along a gridded structured resembling the streets of New York[10]. By separating the painting into each of its colours, and mapping each colour to a different sound parameter, such as pitch, volume or timbre, we can not only see but listen to this rhythm along each of the paths.

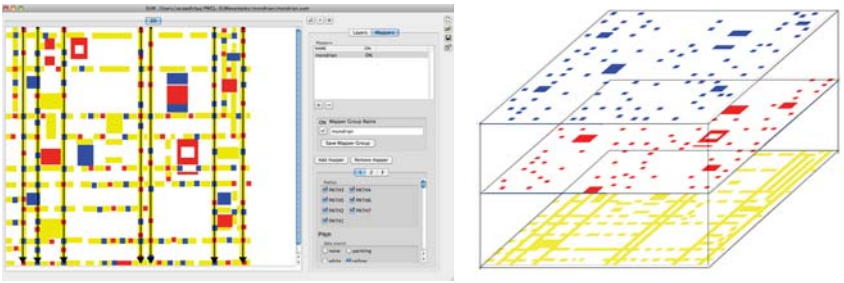


Fig. 5. Sonifying the colour rhythms of Mondrian's *Broadway Boogie Woogie* [10]

Through the sonification of such visual artworks in SUM, we can explore the application of visual composition techniques to musical composition. We can also see the potential for SUM to play any image as an open graphic score. In the following section, we will demonstrate the use of SUM as a tool for computer-aided composition, leading to the generation of a graphic score.

5.2 Computer-aided Composition – Generation of a Graphic Score

The SUM tool, with its vector drawing capability, also supports the creation of graphic scores. The user-defined mapping process means that a composer is free to create his own graphic-sound vocabulary. It supports the creation of a multi-layered graphic score (ie. multiple spatial dimensions), and its playback from any direction, time and speed (ie. multiple temporal dimensions).

As an example, we will show how the graphic score created by Rainer Wehinger for Gyorgy Ligeti's Artikulation, can be generated in SUM and used to explore its playback.

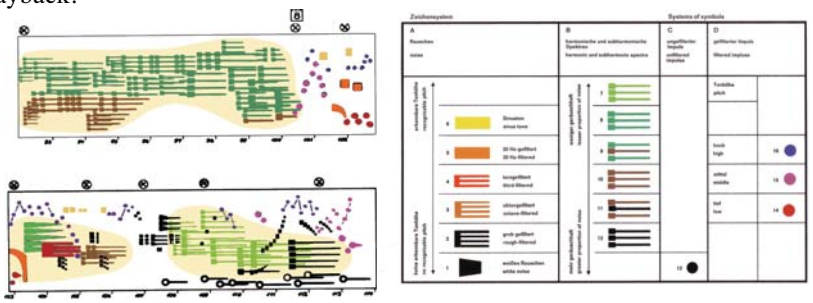


Fig. 6. A section of Wehinger's graphic score for Artikulation (Ligeti) with accompanying colour-coded legend [11]

Wehinger represented each of Ligeti's sound objects graphically, in terms of different forms and colors (see figure 9). As different colors are read as different sound objects in SUM, we can structure our SUM score similarly.

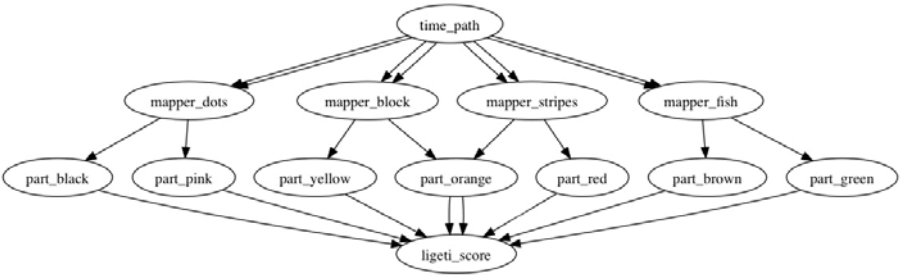


Fig. 7. A possible SUM score structure of Artikulation

The subsequent reading of our SUM score, by any number of user-defined spatio-temporal paths, frees it from its intended linear reading from left-to-right. As seen in figure 8, the same segment of Wehinger's score can be played from different directions and at different speeds.

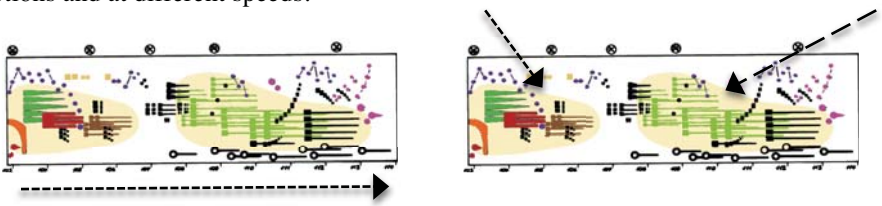


Fig. 8. Different ways of reading Artikulation – linearly as a pianoroll or as an open score

This opens up new possibilities for existing graphic scores to be played in alternative ways and to generate new musical results.

6. Conclusions

As seen above, the structure of the SUM tool supports the integration of image and sound in multiple spatial and temporal dimensions. Growing from the objective to sonify urban maps for a more temporal representation of urban systems, as seen in this paper, we can also use it to compose a multi-dimensional graphical musical score and play it back from numerous perspectives. The flexible structure of SUM allows the audio-visual representation of multiple spatio-temporal relationships in general, from an urban system to a musical score.

Future improvements include the automatisation of the retrieval of the image color palette, and thus the generation of the color-key. We also aim to improve our path-sampling approach in order to more accurately determine the duration of a path.

Acknowledgments.

The authors would like to thank IRCAM and CCRMA for their hospitality during this collaboration. The work of Sara Adhitya has been supported by the John Crampton Scholarship Trustees of Australia. The work of Mika Kuuskankare has been supported by the Academy of Finland (SA137619).

References

1. Adhitya S., Kuuskankare M.: The Sonified Urban Masterplan (SUM) Tool: Sonification For Urban Planning And Design. In: 17th ICAD International Conference on Auditory Display (2011)
2. Kramer, G.: Some organizing principles for representing data with sound. In: Auditory display. Sonification, Audification, and Auditory Interfaces, Addison-Wesley, pp. 185– 221 (1994)
3. Yeo, W. S., Berger, J., Lee, Z.: SonART: A framework for data sonification, visualization and networked multimedia applications. In: 30th ICMC International Computer Music Conference (2004)
4. Laurson M., Kuuskankare M., Norilo V.: An Overview of PWGL, a Visual Programming Environment for Music, In: Computer Music Journal, vol. 33, no.1, pp.19–31 (2009)
5. Kuuskankare M., Laurson M., Expressive Notation Package. , In: Computer Music Journal, 30(4), pp. 67–79 (2006)
6. Baudel, T.: High C draw your music, a graphical composition system inspired by Xenakis' UPIC, http://highc.org/history.html
7. IanniX, a graphical real-time open-source sequencer for digital art, based on Xenakis' works, http://www.iannix.org/en/index.php
8. Unité Polyagogique Informatique du CEMAMu (UPIC), a computerized graphical musical composition tool developed by Iannis Xenakis, Paris, 1977
9. Flanagan, C.: The Bresenham Line-Drawing Algorithm, http://www.cs.helsinki.fi/group/goa/mallinnus/lines/bresenh.html
10. Based on the painting by Piet Mondrian, Broadway Boogie Woogie, MOMA, NY (1942-43)
11. Rainer Wehinger's graphic score (1969) for Gyorgy Ligeti's Artikulation (1958), http://www.tumblr.com/tagged/artikulation

Automatic Interpretation of Chinese Traditional Musical Notation Using Conditional Random Field

Rongfeng Li[1] , Yelei Ding[1] Wenxin Li[1] and Minghui Bi[2],

[1] Key Laboratory of Machine Perception (Ministry of Education), Peking University
[2]School of Arts, Peking University
rongfeng, dingyelei, lwx, biminghui@pku.edu.cn

Abstract. For the majority of Chinese people, Gongchepu, which is the Chinese traditional musical notation, is difficult to understand. Tragically, there are fewer and fewer experts who can read Gongchepu. Our work aims to interpret Gongchepu automatically into western musical notation-staff, which is more easily accepted by the public. The interpretation consists of two parts: pitch interpretation and rhythm interpretation. The pitch interpretation is easily to solve because there is a certain correspondence between the pitch notation of Gongchepu and staff. However, the rhythm notations of Gongchepu cannot be interpreted to the corresponding notations of staff because Gongchepu only denotes ban (strong-beat) and yan (off-beat), and the notations of duration are not taken down. In this paper, we proposed an automatic interpretation model based on Conditional Random Field. Our automatic interpretation method successfully achieves 96.81% precision and 90.59% oov precision on a database of published manually interpretation of Gongchepu.

Keywords: Musical notation, Gongchepu, interpretation, nature language processing, Conditional Random Field

1 Introduction

Chinese poetic songs are noted by *gongchepu*-Chinese traditional musical notation, once popular in ancient China and still used for traditional Chinese musical instruments and Chinese operas nowadays. A *Gongchepu* sample of Chinese poetic songs entitled 天净沙 *Tian-jin-sha* is shown in Figure1.

As illustrated in Figure 1, the melodic notations of *Gongchepu* are noted at the right side of the lyrics, consisted of pitch notation and rhythm notations, which are the two basic characters of a musical notation. Therefore, the interpretation consists two sections, one is pitch interpretation and the other is rhythm interpretation.

This work is supported by the NSFC(No. 60933004).

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

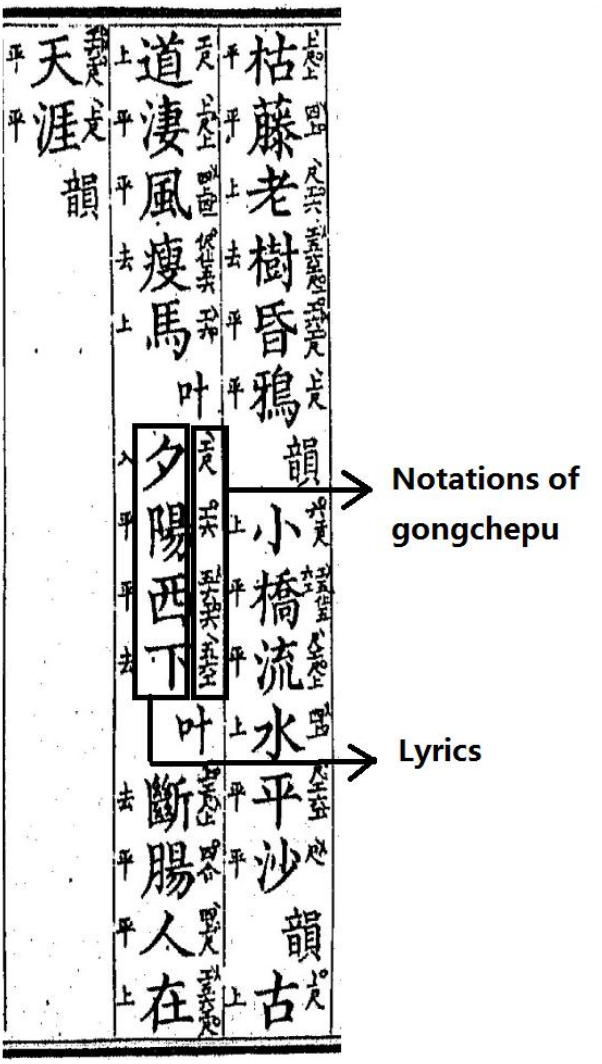


Figure 1. Gongchepu of Tian-jin-sha

For the pitch interpretation, we firstly introduce the details of pitch notations of *gongchepu*. Pitch of each note in *gongchepu* is denoted by 10 Chinese characters:合 *hé*,四 *sì*, 一 *yī*, 上 *shàng*, 尺 *chě*, 工 *gōng*, 凡 *fán*, 六 *liù*, 五 *wǔ*, 乙 *yǐ*. They are equivalent to the notes of solfège system: sol, la, ti, do, re, mi, fa, sol, la, ti. 合 *hé*,四 *sì*, 一 *yī* are pitched an octave lower 六 *liù*, 五 *wǔ*, 乙 *yǐ*. *gongchepu* is named by the character 工 *gōng* and 尺 *chě*.

Once we take 上 *shàng* as the fixed pitch c^1, the range of the 10 characters is g-b^1.

Gongchepu uses the following notations to note other notes in different octaves:

a) Octaves higher: a radical "亻" is added for one octave higher. For example, we use "仕" to represent an octave higher "上". Similarly, the radical "彳" is added to represent two octaves higher.
b) Octaves lower: an attached stroke is added to the ending of stroke of the character to note an octave lower. For example, we use "上" to show an octave lower "上".

Likely, two attached parts are added to represent two octaves lower.

Based on the rule above, the pitch notations of *gongchepu* can be interpreted directly to the corresponding notations of staff.

For the rhythm interpretation, we explain the rhythmic rules of *gongchepu*. *gongchepu* denote the beats by the following notations: The mark "、" represents the stronger-beat which is called *ban*, while the notation "。" represents the off-beat called *yan*. The marks are put at the upper right corner of the first note of a beat. Illustrated from Figure 2 which is written horizontally for convenient reading, we can see the notes separated into beats with the *ban* and *yan*.

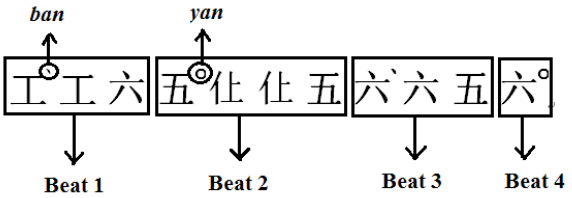


Figure 2. Ban and yan in *gonchepu*

Rhythmic structure of *gongchepu* is formed by the regular combination of *ban* and *yan*. For example, the cycle of 1 *ban* and 1 *yan* forms a 2/4 mater and cycle of 1 *ban* and 3 *yan* forms a 4/4 mater. However, the duration of each note, which should be noted in staff, cannot be specified by the rhythmic mark of *ban* and *yan*. In this case, the rhythm notations cannot be interpreted to the exclusive corresponding notations. For example, if 2 notes are in 1 beat, it can be sung as  ,  or  . If 3 notes are in 1 beat, we could get 4 results:  ,  ,  and  . But whichever should be sung is not restrict by the rhythmic rules of *gongchepu* and can be improvised by the singers. Does this mean that the rhythm in Chinese music is not important as Sachs [1] suggested in his studies of the rhythms of world music? Yang [2] corrects this misconception with the view that in order to perform the music in a proper way, the improvisations should have a certain fixed pattern. In other words, the rhythm of Chinese traditional music does have a certain pattern while the notation of duration of each note cannot be seen in the *gongchepu*.

Despite of all the analysis of the organizational structure of Chinese poetic songs in the past years, almost nothing has been published on the internal rhythmic structure. This is because there are few experts can read *gongchepu* nowadays, and they only teach a small group of students face to face.

In this paper, we proposed a stochastic model to interpret *gongchepu* into staff automatically. Dealing with the rhythm rules of *gongchepu*, the interpretation is similar to part-of-speech tagging in Natural Language Processing. This allows us to use Conditional Random Field to solve the interpretation problem. In recent years, a few musical notation researchers such as Qian [3] and Zhou [4] published their interpretation of the Chinese poetic songs collection, where the *gongchepu* is originally used. We implement our interpretation model on a database their published manually interpretation.

The rest of this paper is structured as follows. We begin with modeling the interpretation problem in section 2. Section 3 introduces the features for the statistical model. Section 4 provides the experimental settings and results. Finally, we draw the conclusion and future discussion in section 5.

2 Automatic Gongchepu Interpretation Model based on Conditional Random Field

In this section, we firstly formulate the interpretations problem. With the formulation, the interpretation problem is transform to a sequence tagging problem which is similar in natural language processing. Then we introduce the most widely used natural language processing model including Hidden Markov Model and Conditional Random Field to solve the interpretation problem.

2.1 Formulations of Rhythm Interpretation

We begin to formulate the interpretation problem by reviewing the rhythm rules of *gongchepu*. The rhythm marks including *ban* and *yan* are put at the upper right corner of the first note of a beat. Thus, notes are separated into beats with the *ban* and *yan*. We denote the beat sequence by $B_1,B_2,\ldots,B_n$ Taking the "Tune of Fresh Flowers" as an example, beats separations are shown in Figure 3.

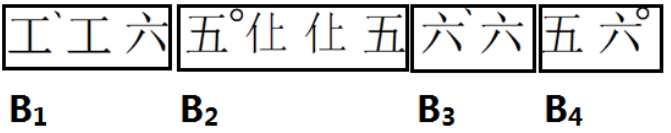


Figure 3. Beat separation by marks of *ban* and *yan*

However, the duration of each note, which should be noted in staff, cannot be specified by the rhythmic mark of *ban* and *yan*. In this case, the rhythm notations cannot be interpreted to the exclusive corresponding notations. For example, if 2 notes are in 1 beat, it can be sung as  ,  or  . We indicate the rhythm pattern of each beat by $R_1, R_2, \ldots, R_n$.

Interpret the notes beat by beat, the interpretation task is illustrated in Figure 4.

In spite of the missing information of the duration of each note, the length of note duration in a beat is relatively fixed. Thus, rhythm patterns of each beat are limited. In

this paper, we conclude 37 patterns p_1, p_2,…,p_{37} which are used in Chinese poetic music. Thus, the value of R_i, i=1, 2, …, n is limited in the patterns set P={p_1, p_2,…,p_{37}}.

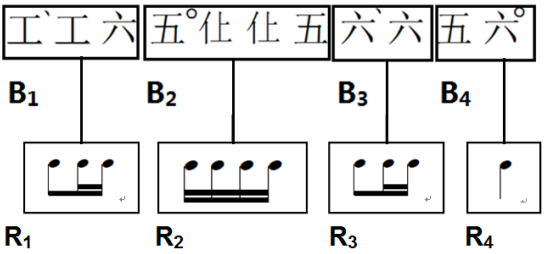


Figure 4. Interpret the rhythm beat by beat

By the above denotations, the interpretation transform to a tagging problem: when the beats sequence {B_1,B_2,…,B_n} is observed, we are required to tag the sequence by the rhythm patterns from a limited set P. This is very similar to the sequence tagging problem in natural language processing.

Once the features $F(B_i)$={$f_1(B_i)$, $f_2(B_i)$,.., $f_m(B_i)$} of each beat are extracted, statistical language processing models such as Conditional Random Field can be applied to the interpretation.

2.2 Hidden Markov Model

HMM is well-understood, versatile and have been successful in handling text-based problem including POS tagging Kupiec[5], named entity recognition (Bikel[6]) and information extraction (Freitag & McCallum[7]). In the rhythm interpretation, the HMM is constructed based on the following assumptions: a) The rhythm pattern sequence { R_1, R_2, … , R_n } forms a Markov Chain; b) The beats B_1,B_2,…,B_n are independent; c) for each rhythm pattern R_i, it only depends on its corresponding beat B_i. The graphical structure of HMM is shown in Figure 5.

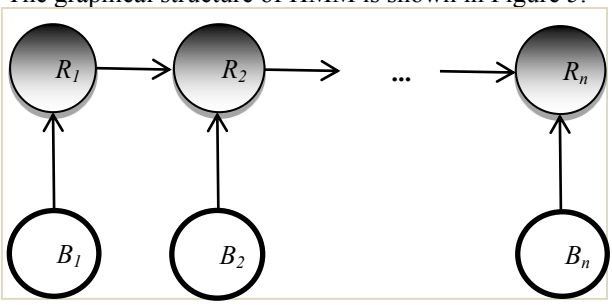


Figure 5. Graphical structure of HMM in rhythm interpretation

2.3 Conditional Random Field

Dealing with the multiple interacting features and long-range dependencies of observation problems, we would be inclined to use Conditional Random Field which is introduced by Lafferty et al [8]. Conditional Random Field have been proven to be efficient in handling different language POS tagging such as Chinese (Hong, Zhang, et al.[9]), Bengali(Ekbal, Haque, et al.[10]) and Tamil(Pandian & Geetha[11]), etc. Compare to HMM, CRF can handle the following undirected graphical structure which is shown in Figure 6.

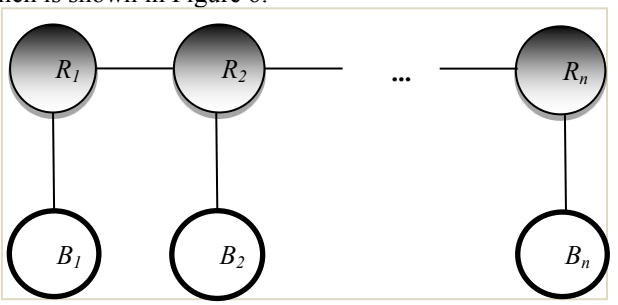


Figure 6. Graphical structure of CRF in rhythm interpretation

Conditional Random Fields are undirected graphic models. Giving an undirected graph G=(V,E). Let C be the set of cliques (fully connected subsets) in the graph. Take the vertex of V as random variable we define the joint distribution of the vertex of V as follows:

$$P(V) = \frac{1}{Z} \prod_{c \in C} \Psi(X_c) \tag{1}$$

Here, X_c is the vertex set of a clique $c \in C$ and Z is the normalizing partition function. Ψ is called a potential function of c. The potential function can be described as the following exponential form:

$$\Psi(X_c) = \exp\left(\sum_i \lambda_i f_i(X_c) \right) \tag{2}$$

In the above model, the undirected graph consists of observations $B_1,B_2,\ldots,B_n$ and states R_1, R_2, … , R_n. Cliques from the above graph consist of two consecutive vertexes which are separated into two classifications: vertex of two consecutive states R_{i-1},R_i and vertex of each states R_i and its corresponding observation B_i. Thus, the exponential form of potential functions can be denoted as the following two functions:

$$\Psi(R_{i-1}, R_i) = \exp\left(\sum_k \lambda_k f_k(R_{i-1}, R_i) \right) \tag{3}$$

and

$$\Psi(R_i, B_i) = \exp\left(\sum_k \mu_k g_k(R_i, B_i)\right) \tag{4}$$

According to the definition of (1), we get the conditional probability distribution:

$$P(R \mid B) = \frac{P(R, B)}{P(B)} = \frac{\frac{1}{Z}\prod_{i=2}^{T}\Psi(R_{i-1}, R_i)\prod_{j=2}^{T}\Psi(R_i, B_i)}{\sum_S\left\{\frac{1}{Z}\prod_{i=2}^{T}\Psi(R_{i-1}, R_i)\prod_{j=2}^{T}\Psi(R_i, B_i)\right\}} \tag{5}$$

Denoting:

$$Z(B) = \sum_S\left\{\frac{1}{Z}\prod_{i=2}^{T}\Psi(R_{i-1}, R_i)\prod_{j=2}^{T}\Psi(R_i, B_i)\right\} \tag{6}$$

(5) can be written as:

$$P(R \mid B) = \frac{1}{Z(B)}\exp\left\{\sum_i\sum_k \lambda_k f_k(R_{i-1}, R_i) + \mu_k g_k(R_i, B_i)\right\} \tag{7}$$

Here f_k is the feature function and g_k is the state feature functions. $\lambda_1, \lambda_2, \ldots, \lambda_T, \mu_1, \mu_2, \ldots, \mu_T$ are parameters to be estimated from training data.

To apply the above models, we should extract the features of each beat, which are discussed in the following section.

3 Feature Selection for Automatic Interpretation

Wise choice of the features is always vital to the performance of the statistical models. Chinese traditional music does not have harmony, polyphony, or texture. Thus, we only concern about the melody and select the proper features based on the opinions of the Chinese opera performance as follows.

- **Notes Sequence (NS):** The higher and lower octave symbols expand the 10 characters in *gongchepu* into 38 characters. Encoding these characters, we can get the original text features of the notes sequences.
- **Numbers of the Notes(NN):** Sequence of the notes numbers forms the approximately rhythmic structure. Rhythmic pattern is usually related to the notes number of previous beat. In the example of "Tune of Fresh Flowers" in figure 5, we consider the third beat "六`六五" which is a three-note beat and the previous beat has four notes. Therefore, it preferred to determine the rhythmic pattern as  rather than  to avoid a too compact rhythmic structure.
- **Pitch Interval Direction and Position(PIDP):** The concept of "interval direction and position" is introduced by Williams(1997) for melodic analysis. Williams use "+" for rising direction of the pitch interval and "-" for the falling direction. Moreover, pitch interval is measured by

chromatic scale. For example, the pitch interval direction and position of the section of "Tune of Fresh Flowers" is illustrated in Figure 7.

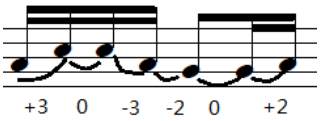


Figure 7.Pitch interval direction and position of "Tune of Fresh Flowers"

4 Experimental Result

The experiments of *gongchepu* interpretation were based on the *gongchepu* of *Sui-jin-Ci-pu* collected by Xie[12] which collected poetic songs of Tang, Song and Yuan Dynasties of ancient China. *Sui-jin-Ci-pu* collected over 800 songs, but only a few of them have been interpreted. We trained our statistical models based on Qian [5]'s manually interpretation. We selected 60 songs from the 96 of Qian's interpretation to set up our database. The database included 969 melody segments and amounted to 6347 beats. According to the different number of notes within a beat, the beats were separated into 6 types. The dataset was randomly divided into two parts with similar distribution of different types of beats. 3174 beats were used as training data while the left 3173 were reserved for test.

Table 1: Data size of *gongchepu*

Numbers of notes with in a beat	**Training data size**	**Testing data size**	**Total data size**
1	1187	1017	2204
2	1110	1322	2432
3	647	676	1323
4	210	152	362
5	19	5	24
6	1	1	2
Total	3174	3173	6347

Table 1 shows the data size of the *gongchepu* for training and testing. In the table, we can see there are only 24 beats with 5 notes and 2 beats with 6 notes. 99.59% of beats in the dataset have more than 4 notes.

Two method Hidden Markov Model (HMM) and Conditional Random Field (CRF) which were introduced in Section 2 are applied using three single features: notes sequence (NS), numbers of notes (NN), pitch interval position and direction (PIDP) and their combinations: NS+NN, NN+PIDP, NS+PIDP, NS+NN+PIDP. The experimental results of interpretation precision and oov precisions are shown in Table 2.

Table 2. Interpretation precision and oov precisions

	precision		oov precision	
Features	HMM	CRF	HMM	CRF
NS	84.34%	87.86%	47.85%	67.62%
NN	83.43%	85.55%	68.43%	78.84%
PIDP	84.82%	85.97%	57.92%	77.53%
NS+NN	85.64%	89.67%	75.67%	80.23%
NN+PIDP	86.74%	89.56%	77.28%	81.55%
NS+PIDP	85.49%	89.89%	76.42%	79.88%
NS+NN+PIDP	87.38%	90.05%	78.27%	82.03%

The results from table 2 shows that CRF get better performance than HMM and achieve 90.05% precision and 82.03% oov precisions using the combination feature of NS+NN+PIDP.

We analyzed the oov beat and found that most interpretation error occurred in handling the beats which have 3 notes. For example,  is always misinterpreted into  .

After rhythmic pattern tagging, we can interpret *gongchepu* automatically. The interpreted staff of the *gongchepu* of 天净沙 Tian-jin-sha in Figure 1 is shown in Figure 8.

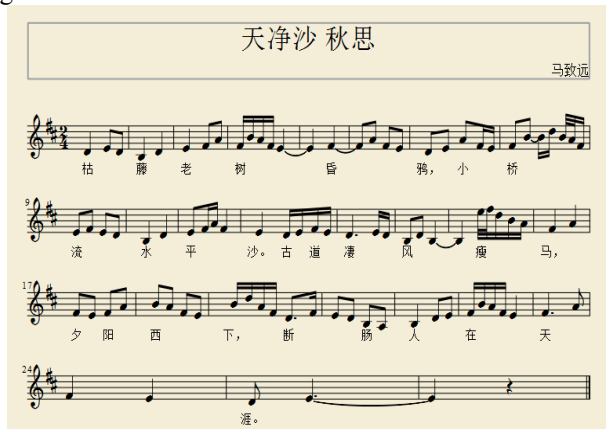


Figure 8.Interpretation of Tian-jin-sha

5 Conclusions and Future Discussions

This paper proposed an automatic interpretation of *gongchepu*. We apply Hidden Markov Model and Conditional Random Field to solve the interpretation problem. Three single features: notes sequence (NS), numbers of notes (NN), pitch interval position and direction (PIDP) and their combinations: NS+NN, NN+PIDP, NS+PIDP, NS+NN+PIDP are selected for the interpretation model.

Experimental results showed that the precision of interpretation by CRF achieved 90.05% and the oov precision was 82.03%. It will be very helpful for reading and singing the Chinese poetic songs noted in *gongchepu*. Furthermore, our work will have positive influence on the protection of the ancient Chinese traditional culture, for the number of the experts who are able to read *gongchepu* is decreasing and the way of singing Chinese traditional poetic songs will most likely fade in the following generations.

Obviously, the sample size of the *gongchepu* database (6347 beats) is much smaller than the corpus in NLP. However, music is more abstract than natural language, and music is an easier way for listener to understand and accept, while natural language may cause many unpredictable misunderstandings. Thus our work, training on the musical notation database, which is much smaller than the NLP corpus, is still credible.

Melodic features only bring a superficial knowledge in understanding the rhythm of *gongchepu*. Actually, Chinese language plays an important role in the development of Chinese music. Thus in the further research, we will take the linguistic features in consideration.

References

1. Curt Sachs: Chinese Tune-Title Lyrics. The Rise of Music in the Ancient World. London (1943)
2. Yinliu Yang: Gongchepu-qian-shuo "Introduction of gongchepu". Renmin yinyue chubanshe. Beijing (1962)
3. Rengkang Qian: Qing-jun-shi-chang-qian-chao-qu "Interpretation of Suijin cipu". Shanghai yinyue chubanshe, Shanghai(2006)
4. Xuehua Zhou: Nashu-ying-qu-pu-jian-pu-ban "Interpretation of nashu". Shanghai jiaoyu chubanshe. Shanghai (2008)
5. Julian Kupiec: Robust part-of-speech tagging using a hidden Markov model. Computer Speech and Language, 6, 225–242. (1992)
6. Daniel M.Bikel, Richard Schwartz, & Ralph M.Weischedel: An Algorithm that Learns what's in a name. Machine Learning Journal, 34, 211–231. (1999)
7. Dayne Freitag & Andrew McCallum: Information Extraction Using HMMs and Shrinkage. In Papers from the AAAI-99 Workshop on Machine Learning for Information Extration, pp. 31–36 Menlo Park, California. AAAI. (1999)
8. John Lafferty, Andrew McCallum and Fernando Pereira: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the Eighteenth International Conference on Machine Learning. (2001)
9. Mingcai Hong, Kuo Zhang, Jie Tang & Zijuan Li: A Chinese Part-of-speech Tagging Approach Using Conditional Random Fields. Computer Science, Vol. 33, No. 10, pp. 148-152. (2006)
10. Ekbal Asif, Rejwanul Haque, and Sivaji Bandyopadhyay: Bengali Part of Speech Tagging using Conditional Random Field. In Proceedings of Seventh Inter-national Symposium on Natural Language Processing. Thailand (2007)
11. S. Lakshmana Pandian, T. V. Geetha: CRF Models for Tamil Part of Speech Tagging and Chunking. Proceedings of the 22nd International Conference on Computer Processing of Oriental Languages, 11-22. 42. (2009)
12. Yuanhuai Xie: Sui-jin-ci-pu "A Collection of Song". (1844)

Music Dramaturgy and Human Reactions: Music as a Means for Communication

Javier Alejandro Garavaglia,

Sir John Cass Faculty of Art, Media and Design - London Metropolitan University
41-71 Commercial Road, E1 1LA – London – UK
j.garavaglia@londonmet.ac.uk

Abstract. The main topic of this paper refers to how music communicates and to what it communicates, either considering or not the usage of modern technologies. Based on the categorisation of music dramaturgy proposed in one of his pasts articles [1][2], the author sets the main focus on what happens in the mind of listeners (perception) during a performance (and afterwards) of music rather than considering only the perspective of the creator (intention). Thus, the article not only connects the fields of neuroscience with that of semiotics, but also is a reflection from a philosophical perspective of how the dramaturgy of music affects the perception by arousing reactions (emotions and thoughts) in the audience.

Keywords: Dramaturgy of Music; Music semiotics; Neuroscience; Prototype Theory; Exemplar Theory; Multiple-Trace Memory model; Categorisation.

1 Introduction

The subject of music dramaturgy has been treated across time in different ways and from different perspectives; in the last two decades Landy and later Weale have performed a fundamental research in the field [3][4][5]. The research presented in this article, although related to Landy and Weale, focuses on music in a general and broader sense. Fundamentally, the research I have carried out so far [1][2] includes questions seeking for the clarification of, for example, how the relationship creator-listener works in musical situations or, what happens in the mind of the listener whilst perceiving a piece of music. The research presented herewith is therefore a further development of the classification of music dramaturgy presented in my article *Music and Technology: What Impact Does Technology Have on the Dramaturgy of Music?* [1]. Figure 1 summarises the complete typification of music dramaturgy proposed therein. For the current article though, the main subject focuses specifically on the relationship between music and human reactions, giving special attention to how the human brain reacts to musical stimuli.

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

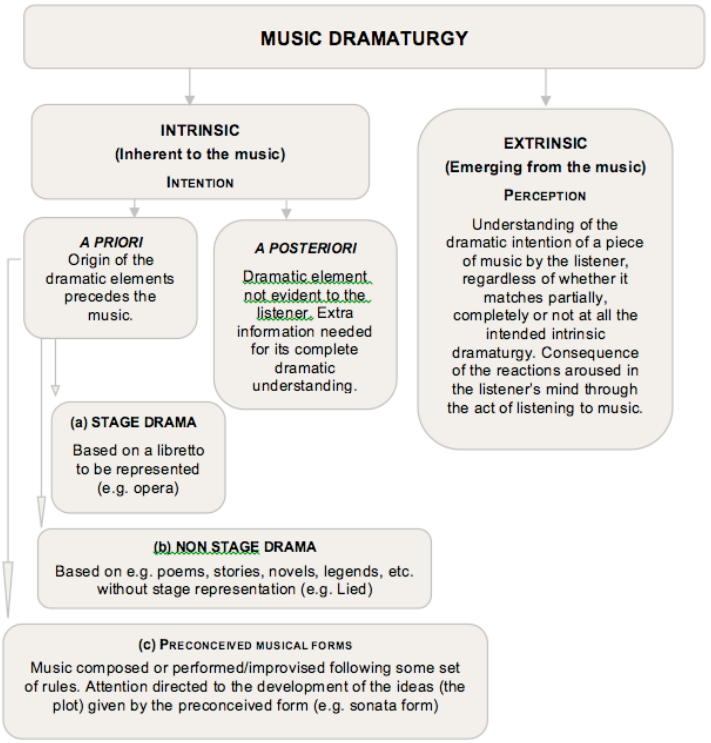


Fig. 1. Music dramaturgy: different types and subcategories.

2 Music as a Means for Communication: Musical Discourse and Human Reactions in a Musical Communication Chain

To begin with, it is important to define exactly what is meant with music and music dramaturgy in this context. Even though the concept of music has been defined and redefined across time, the following definition by Levitin is not only rather comprehensive and clear, but it also addresses my fundamental concerns as a composer, especially considering how human perception regards some sounds as musical or not:

> The difference between music and a random or disordered set of sounds has to do with the way these fundamental attributes combine, and the relations that form between them. When these basic elements combine and form relationships with one another in a meaningful way, they give rise to higher-order concepts such as meter, key, melody, and harmony. [6]

In [1], I have already defined the dramaturgy of music as:

> As we can see, the word 'dramaturgy' has its origin in the German word *Dramaturgie* and its roots can be found in the ancient Greek word *dramatourgia*. However, the main term to consider should be *drama*: its meaning is always related to the concepts of 'action' or 'event'. Aristotle, in chapter 3 of his *On the art of poetry*, describes drama as something 'being done'. The word dramaturgy implies the actual composition or

'arrangement into specific proportion or relation and especially into artistic form' as well as the knowledge of the rules for gathering these concepts onto a (normally) known and preconceived structure (originally, the Greek tragedy was meant hereby).

Ultimately, we can define the dramaturgy of music as the way in which the creator and the listener represent in their minds the flow of a musical occurrence (that is the development of one sonic-event coming from a previous one and leading to the next), which constitutes an entity (ontologically) that as such is unique in itself, as might also be its mental representation (psychologically); however, both cases of 'uniqueness' might not be most of the time quite the same, as we shall see later. The series of sounds organised according to the rules of each and every musical 'being' (the word 'being' is here used ontologically, meaning anything that can be said to *be* immanently, as not always might we refer to a composition when confronted to music-listening, mostly if we consider music from outside the western culture), are the events involving an 'interesting or intense conflict of forces', as seen above in one of the definitions of dramaturgy. And, as in the case of the original meaning of the word in ancient Greece, these forces do happen during a performance. The forces in place are the emotions/thoughts aroused by the sounds of the performance, which produce a mental representation of what is occurring in the piece of music: its emergent dramaturgy.

From this definition, we can infer, that the subject of human emotions is core to the field of music dramaturgy. Following the definition of music given above, the 'basic elements [that] combine and form relationships with one another in a meaningful way' are those which need to be communicated in a chain, so the next step is to present the communication process in a musical situation and all of the elements taking part in it. The following subsections give an explanation of the concepts of musical discourse, musical communication chain and human reactions.

2.1 Dramaturgy in the Musical Discourse: The Communication Process

Any imaginable type of music is capable of awaking in the listener reactions such as thoughts (i.e. mental representations of the sonic events and their subjective meaning) and emotions, all of which may or may not be in tune to the original intention of the creator of that particular music. Reception of music dramaturgy can only be possible if a communicative process is established. This process requires three elements for its existence: (a) actors involved in the communication process; (b) medium in which the dramaturgy will be carried; (c) human reactions.

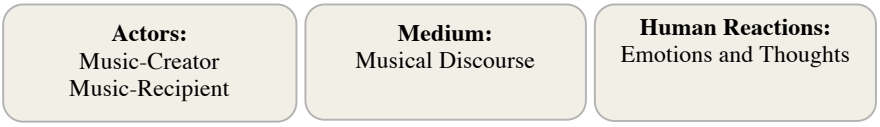


Fig. 2. Communication process in music: minimum elements required.

Hence, in order for music to be in a position to express something, a communicative process must be established. In this way the creator of a certain type of music (generally, but not exclusively, the composer), delivers through a process (the musical performance, meant here in a broad and generic way), a musical discourse containing the main intentions, which will be finally perceived by a human-recipient (generally, an audience of listeners). In this communication chain, the reception may or may not equal the original intention; moreover, the perception of the musical discourse can even result (as explained in [1]), in a rather opposite

understanding of the original intention conceived by the music-creator.

Whichever the response of listeners to music may be, this response is generally called *arousal* in psychology, which is defined as 'to rouse or stimulate to action or to psychological readiness for activity' [7].[1] According to this, the act of perception should produce in the listener diverse reactions, which can mainly be circumscribed to emotions and further thoughts or reflection of what has been listened to. Arousal is sometimes also referred to as activation [7].[2] A communication chain emerges from this concept as represented in Fig. 3:

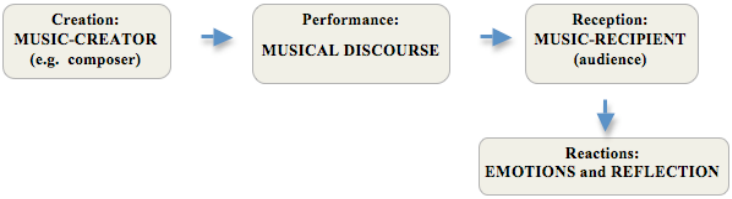


Fig. 3. Music's communication chain (first stage).

If the listener is not in a position to experience any reaction at all, this will imply, that either the event being listened to contained no message at all (i.e. there is no musical discourse present) or the listener is not in a position to understand the musical discourse as such. In the first case, the absence of arousal is due to an objective failure in the chain, as the object missing is outside the mind of the listener. In the second, on the other hand, the absence of arousal is due to a subjective failure in the chain, as the musical discourse exists, but cannot be understood by listeners due to diverse causes such as, for example, cultural background. This paragraph by Berio clarifies the matter further, mostly at the end:

> Music must be capable of educating people to discover and create relations between different elements (as Dante said in the Convivio, 'music is all relative'), and in doing that it speaks of the history of man and of his musical resources in all their acoustic, and expressive aspects. I'm interested by music that creates and develops relations between very distant points, and pursues a very wide transformational trajectory (....). The listener has to be aware that there are different ways of grasping the sense of that trajectory (....).[3]

If there is no arousal (a complete absence of any reaction), regardless of which of the two cases mentioned above is considered, the result will undoubtedly be a complete failure at the very core of the basic communication principle. If music should contain and express a certain type of dramaturgy [1], the first case should not be possible, as the musical discourse must be indeed present at every musical

[1] Encyclopaedia Britannica Library - 2004: Arouse [7].

[2] Activation: also called arousal in psychology, the stimulation of the cerebral cortex into a state of general wakefulness or attention. Activation proceeds from various portions of the brain, but mainly from the reticular formation, the nerve network in the midbrain that monitors ingoing and outgoing sensory and motor impulses. Activation, however, is not the same as direct cortical stimulation by specific sense receptors, such as being awakened by noises. It involves, rather, a complex of impulses that are both internal and external to the body. (Encyclopaedia Britannica Library - 2004: Activation) [7]. See also Chapter 9 of [8], written by Simonton, which deals with the subject too.

[3] Luciano Berio: Two Interviews with Rossana Dalmonte and Balint Andras Varga [11].

manifestation, regardless of whether understood or not by the listener. The second case however, does happen and rather often; this is mostly due to diversity of the cultural backgrounds of different listener types. But, paraphrasing Berio, if people can be 'educated' in this sense, this case may only be circumstantial and not final.

Having said that, in the case in which the musical discourse is both present and understood as such by the listener, this implies the presence of a (musical) communication process and therefore, by reacting to these stimuli, listeners can connect an external musical discourse to their own interior and personal world (or *phaneron*, to use Peirce's terminology [9]) The next step will be their own understanding of the event. Landy, based on Nattiez adds the following:

> Nattiez has offered a useful definition of meaning for an individual apprehending that object, as soon as the individual places the object in relation of his [or her] lived experience-that is, in relation to a collection of other objects that belong to his or her experience of the world. [4]

The music-creator is who exposes the music work openly from the inside to the outside, as it is only outside the self that any work can be contemplated, regardless of whether by other listeners or by the him/herself.[4] As music is a temporal act *per se* (it happens *during* time), it can be inferred that a musical discourse cannot happen without the following two dimensions: space (the outside world) and time. The contemplation of a piece of music will happen inside each 'music-recipient' in a physical space during a determined lapse of time. It is through this contemplation that the dramaturgy of the musical discourse may become apparent. This implies that the recipient has to be acquainted with the type of musical discourse listened to, which brings us to the subjects of cultural background, expectation and mental contours[5]. As the brain adapts itself in a very early stage in life (as early as inside the womb), it stores information of the surrounding world in the long term memory, what helps later in life to recall well known contours (e.g. in music: harmony, melody, rhythm, etc.). This leads to expect due to previous knowledge similar results in new, never experienced before but yet similar musical contours. [6] The general cultural background of each individual will have similar results in how to imagine the music heard by relating to already learned contours. If the models or contours are known to the listener, the brain can predict and even be predisposed to understand the dramaturgy of a given music by comparing it with previous experiences. Cognitive science describes this as a mental schema: a framework within which the brain places (stores) standard situations, extracting those elements common to multiple experiences [6]. In music appreciation, familiarity (what creates the network of neurons in the brain forming the according mental schema) brings the listener's attention onto music styles that the brain may or may not recognise. Even if the listener will generally not be familiar to every piece of music listened, those mental schemas may guide the brain to form new neural connections to recognise new elements with which it is, partially or totally, not familiar. This expectation can be broken with surprise if new elements appear (elements unknown to the listener's brain), and depending on how they are combined in a piece of music, the schemas coming out of this appreciation may be stored in the brain and be recognised in future auditions of the same piece or even others, which share similar characteristics.

Therefore, expectation plays a crucial role in whether recognising or not what is

---

[4] The creator can also be the end-recipient, when it comes to the reproduction of own music.

[5] Contour: 'the general form or structure of something'. Term used also, to determine some 'meaningful change in intonation in speech'. [10]

being listened to. Hence, it suffices to be in such a position as to perceive the musical discourse as a musical event and not as a mere conglomerate of sounds without any connection between them. The listener is required only to possess some basic information (mainly through expectation, regardless of its degree), which will enable him to recognise that he is being confronted with a musical event and not with something else. Here, the listener's cultural background plays an eminent role. The concept of what music is has changed through the passing of time; however not only time is of vital importance here, but also where (referring to style and culture) the music may have originated. Nonetheless, for listeners to understand a musical discourse capable of arousing emotions and reflections, all of which will develop a dramaturgy in their minds, it is a prerequisite for them to understand that what is being perceived is music and nothing else. Thus, the logical consequence is that this type of dramaturgy on the listeners' side is a subjective occurrence inside their minds originated in the act of listening. This, in spite of such a representation having its origin in an external source, the music itself, which, through the musical discourse carries an inner expressive intention given subjectively by its creator. This subjective intention though, does not need to be apparent (and in many cases, it may well not be) and is in many cases unknown to the listener. Berio's following statement sheds some light on the matter:

> My listener will have the possibility to understand the music in different ways: in a way, if he succeeds in deciphering the references; in another way, if he is not familiar with them.[6]

Following this idea (similar to Weale's concept of intention/reception [5]) and applying my personal reflections both as composer and music-listener, is that I proposed in [1] two main categories of music dramaturgy, which were summarised in figure 1 above. Thus, these categories two can be defined as:

- Intrinsic or inner music dramaturgy: the inherent message that the musical discourse carries within itself, which can be identified during the time of conception of any type of music, in which the creator models his intentions into a musical discourse.

- Extrinsic or emergent music dramaturgy: which is activated in the recipient's mind by the act of listening. This dramaturgy arises only through the contemplation of music and may or may not be the same dramaturgy carried by the music being listened to (the one intended by the creator). It becomes apparent only after human reactions have been aroused in the listener's mind. Figure 4 shows a complete chart of the communication chain, as explained in this section.

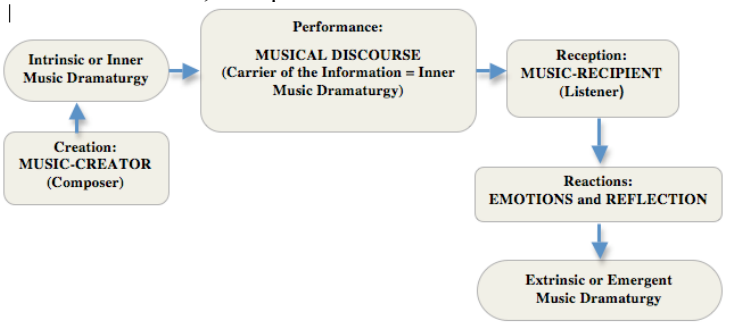


Fig. 4. Music's communication chain (complete chart).

[6] Ivanka Stoianova. Luciano Berio. Chemins en Musique, Paris 1985 [12].

As already mentioned, music, opposed to some other types of art, depends for its existence on an external and objective time: it is a 'process'. The communication chart (Fig. 4) represents the process in which music happens, and it requires time to exist. Time however, may imply here a double connotation: there is an actual time, in which the performance of music occurs[7], and yet another, from the perspective of the listener, which must be regarded as a relative value due to the subjectiveness of the situation. This concept is linked to the philosophy of Henri-Louis Bergson. According to Bergson's (referring to time), *duration* is:

> [T]he development of a thought that gradually changes as it takes shape.... Time is invention or it is nothing at all. [13][14]

Moreover:

> For … the philosopher, time is a free-flowing medium that depends for its perception on what is filling it. In Time and Free Will (1889), Bergson said that time could not be evenly divided as by a clock, whose measurement dissolves time into tiny points in space. [14]

Further, to this, the concept of *event* cannot be ignored. An *event* is also a process and can refer to many fields, including those of philosophy. The following definition states that:

> Broadly understood, events are things that happen—things such as births and deaths, thunder and lightening, explosions, weddings, hiccups and hand-waves, dances, smiles, walks. Whether such things form a genuine metaphysical category is a question that has attracted the sustained interest of philosophers, especially in the second half of the 20th century. [15]

In the field of Philosophy, a definitive and unique definition of *event* does not yet exist, and multiple theories co-exist. According to Kim [16], events are comprised of three elements: object {x}, property {P} and time {t}; by combining them using the operation {x,P,t} Kim states that events are defined. From the point of view of perception, the structure of an event must be discerned by recipients, who will save in their memory a certain amount of information about the contemplated event (depending on factors such as attention, concentration, previous experiences, etc). The structure of the event can be seen as its 'dramaturgy'. If Kim's theory is transferred to music, then a musical event can be regarded as a continuum across time of different combinations of sounds-and-breaks {x} gathered with a particular purpose {P}, at a particular time {t} The purpose {P} has a meaning, which may substantially change if, for example, an alteration in the order of events occurs. Thus, and even though the constitutive elements may still the same, their order in time is different, with an impact on the manner that music may be perceived and understood, and therefore, different human reactions arouse.

The following subsections explain in detail the concepts of musical discourse and of human reactions, including emotions.

2.2 Musical Discourse

[7] I refer hereby to the actual time outside our perception crafts. In the case of other types of art, such as plastics, even though a painting may induce some kind of dramaturgy, the painting in itself is ontologically timeless: the time of contemplation depends exclusively on a subjective act from the side of the recipient (for example, how long the recipient will be watching at it). In the case of music, there is an actual, objective time, determined by the duration of each performance.

Definitions of communication, such as 'an act or instance of transmitting' [10], include the concept of transmission. To transmit is usually defined as 'to send or convey from one person or place to another' [10]. In both cases, it is implicit that there is 'something' being transmitted. Caesar explains [17] that, in *A Theory of Semiotics*, Eco gives a special consideration to the relationship between the words 'communication' and 'signification'. Here, even though both concepts are different in their meaning, they are 'not mutually excluded' in the field of semiotics [17]. Further, Caesar makes clear, that the distinction on which Eco bases this concept relies on the fact that:

> …[S]emiotics is coexistence with signification which occurs only when the communicative act envisages a potential human addressee acting as an interpreter of the message (and not a receiver merely responding to a stimulus). [17]

This implies that a semiotic of signification can exist without a semiotic of communication, but not inversely. Therefore, Caesar deduces that Eco includes a human factor in the chain, given the fact that signification, according to Eco, cannot occur if there is no human addressee interpreting the message. It is in this sense that my views about music dramaturgy and its communication chain are presented here: my main semiotic interests in music composition are its semantic[8] and essentially, its pragmatic[9]values, rather than its syntax[10].

In music, the message interpreted by the addressee in the communicative act is the musical discourse, the main object of transmission in a musical situation. To discuss the musical discourse, it is of advantage to look at the definition of both words first.

Discourse has several definitions, depending on the usage of the word. Related to music, these two definitions may be the closest: 'formal and orderly and usually extended expression of thought on a subject' or 'a mode of organizing knowledge, ideas, or experience that is rooted in language and its concrete contexts' [10].

The expression of thought or the organisation of ideas rooted in language exists through signs and symbols, which confer to the discourse its syntax and semantic aspects. This means, that by referring to musical discourse, we enter the domain of musical semiotics. Nattiez's writings are arguably the main source to look for the concepts of musical discourse and musical semiotics.

> In contradistinction to human language, musical discourse does not strive to convey clear, logically articulated messages. For this reason, we may well ask whether one can speak of such things as "musical narrativity". … Musical discourse inscribes itself in time. It is comprised of repetitions, recollections, preparations, expectations, and

---

[8] Semantics is basically the relationship between signs and what they refer to. 'The word "semantics" itself denotes a range of ideas, from the popular to the highly technical. It is often used in ordinary language to denote a problem of understanding that comes down to word selection or connotation.' (http://en.wikipedia.org/wiki/Semantics) [18]

[9] Pragmatics is the relationship between signs and their impact on those using them. 'Studies how the transmission of meaning depends not only on the linguistic knowledge (for example, grammar, lexicon, etc.) of the speaker and listener, but also on the context of the utterance, knowledge about the status of those involved, the inferred intent of the speaker, and so on.' (http://en.wikipedia.org/wiki/Pragmatics) [18]

[10] Syntactics refers to the relationship between signs in formal structures. 'The study of the principles and rules for constructing sentences in natural languages. In addition to referring to the discipline, the term syntax is also used to refer directly to the rules and principles that govern the sentence structure of any individual language...'.
(http://en.wikipedia.org/wiki/Syntax) [18]

resolutions, and in the realm of melodic syntax. [19]

Applied specifically to music, the discourse should therefore be the means of carrying the musical expression. However, the expression intended by the music-creator may or may not be understood by the listener as conceived, depending on each particular case. However, without a musical discourse it is impossible to establish the required communication act, as there would not be any element (message) to be communicated. This does not imply though, that a music discourse will make a particular composition more accessible or even will determine a unique and universal view to that particular piece of music (thus, determining the pragmatic level of its semiotic contents). On the contrary, this means that, on the one hand, each listener will understand the same composition and/or performance differently from others (with a wide degree of variation among them); on the other hand, as this understanding is absolutely tied to the cultural environment and personal background of each particular subject, it may not connect at all with the intention of the composer/creator. Nattiez says:

> If the listener, in listening to music, experiences the suasions of what I would like to call the narrative impulse, this is because he or she hears (on the level of strictly musical discourse) recollections, expectations, and resolutions, but does not know what is expected, what resolved. The listener will be seized by a desire to complete, in words, what music does not say, because music is incapable of saying it. Such things are not in music's semiological nature. [19]

Yet, the clearer the musical discourse of a piece of music (in its syntactic and semantic dimension), the better the reception that may be obtained from the original intention assigned to that music. In any case, communication (regardless of the level and degree of its understanding) has been established when a musical discourse is present, provided it can be understood as such. The understanding of it presupposes therefore the existence of Eco's 'addressee' [17].

Music has been contemplated in the past from rather diverse angles. Kivy [20] explains how to interpret Aristotle's definition of the Greek word **μνμησιζ** (*mimesis*, meaning imitation), when applied specifically to music. He starts by quoting Thomas Twinning's interpretation of this word in his 1789 translation of Aristotle's *Treatise on Poetry*. Kivy agrees with Twinning, that the word imitation should be understood as what was actually meant in its own time, closer to 'expression' rather than the modern concept of 'imitation'. As Kivy quotes from the Shorter Oxford English Dictionary, the definition for the word imitation is to 'copy' or to 'reproduce' and, moreover, it is a 'counterfeit' or an 'artificial likeness'. Associated with this, Kivy also explains Aristotle's claim in his *Politics VII* (1340a): here, what music does imitate is in fact 'emotions and states of human character'. [20]

2.3 Human Reactions

Listeners react to the musical discourse in distinctive ways; these reactions include emotions. Unfortunately, the word 'emotion', which is generally associated with feelings, thoughts and behaviours, is quite ambiguous in meaning, and depending on which line of research is followed, a different understanding of the very concept of emotion will arise. Some lines of research have made synonyms of the words emotion and feeling. Furthermore, there seems to be neither an established procedure nor an agreement in the research community so far to define the number

and nature of a standard set of different categories of emotions [21]. For the purpose of this article however, I shall treat hereby emotions and thoughts as separate entities.

In spite of this ambiguity, listeners' reactions, specifically emotions, need further analysis because, even if emotions may be present in the majority of cases throughout the entire listening process (and, most likely, also beyond), they can be either the first reaction to the act of listening (previous to any rationalisation) or the reaction to some reflection about what has been listened to. This distinction can affect the entire communication process, and therefore, the perception of the musical discourse. Music is listened to at the very first stage through the senses (mainly through the sense of hearing), and this first reception arouses almost immediately in the listener some type of reaction, in many cases, an emotional reaction, which can be extremely variable depending on each particular situation.

Some musical materials such as chords and melodies in western tonal music tend to produce some common emotional reactions in (at least) western audiences: just as an example, minor chords or even tonalities seem generally to be associated with a sad or melancholic mood, arousing a similar type of reaction. A piece of music, however, is a complex combination of different musical materials, such as chords, harmonies, melodies and even layers of sound. From the perspective of music semiotics, these elements isolated constitute the syntactic 'signs' of the musical narrative. When gathered together, the tension created by those elements is what may produce the understanding of and reflection on what has been listened to; after that, an emotional reaction may follow, which may not be the same as it could have been for particular elements of that piece (such as isolated chords, melodies, etc.), but an emotional reaction that arises from listening to the entire work. In other cases however, that tension may resolve directly in emotions, which then may influence the ulterior understanding of a work of music and are therefore prior to any reflection or thought. Hence, two situations can be distinguished, which I regard hereby as first and second cases of arousal in music perception:

- reflection/thoughts ➔ emotions (first case) and
- emotions ➔ reflection/thoughts (second case)

Both cases relate to the empathetic listening behaviour by Delalande [22][1].

Sometimes, however, the arousal of emotions in the listener's mind may not happen after reflection. In this particular situation, the dramaturgy that emerges is solely the consequence of reflection. The opposite however (no reflections after the emotions), is indeed rare, as our brain has evolved in such a way that it is programmed to imagine stories, thus reflecting on what it experiences. Muller [23] refers to the research in the 1970s and 1980s by Roger G. Schank, who examined the issue of how human beings think and further, how those thinking processes influence our behaviour; through this research, Schank attempted to develop artificial intelligence programmes for computers. This research concluded with the idea, that the human brain is programmed to think in terms of stories. Quoting from Muller's article:

> A human brain may receive thousands of pieces of information daily. Most of it we can't retrieve, even minutes later, while other information can stay with us for years, and we can easily recall it. Why? Because the information that we tend to remember is presented in the context of a story about the information, person, or event. [23].

In the cases in which emotions happen after reflection, they can however vary with the audition at different moments or situations (for example in a different mood) of the same piece of music (thus, with the same musical discourse). In these cases,

emotions can even induce the listener to a different understanding. Therefore, emotions can either be the consequence of the reflection on what was heard or the trigger to an interpretation. It mainly depends on the personal background and state of mind of each listener for one case or the other to happen. Furthermore, in the second case, thoughts can trigger further emotions, which may vary in some degree the former understanding, and changing it accordingly. The chain can go indefinitely. Thus, emotions can be two-fold, as they may predispose listeners to understand the music in a particular way by defining or at least influencing how, the musical intention can be perceived or they may be the result of that understanding. The understanding that emerges in the recipient's mind can change from time to time depending on moods, cultural background, experiences of life, expectation, and so on, producing different reactions in the same person at different times, even in cases in which, the same piece of music (even in the same interpretation, or same recording) is being listened to.

Research in the area of emotional reactions to music situations shows that in the last hundred years it has concentrated mainly (in some cases even exclusively) on the parameter melody. This can be observed in several cases, such as the writings by Budd [24] or Cohen (in [8], Chapter 11), even though the latter includes film music from a perspective that does not treat solely the melodic aspect. However, developments in music since the Italian Futurism in 1909, where other musical parameters rather than melody constitute the essence of some music, seem to be rather ignored or left aside. I refer here to cases such as electronic, acousmatic and interactive music: all these types not only work mainly based on concrete sounds and noise, but quite often their most likely constitutive musical parameters are timbre or sound spatialisation. Moreover, my disagreement with these analysis on music and emotions (such as Simonton's in [8], Chapter 9) relies on the fact, that they do not only focus on the essence of the emotional reaction over melodic aspects of the music alone (ignoring other music parameters, such as harmonic tensions or timbre), but also, that this view implies to put the weight of the reaction on the music rather than on the listener. According to this view, syntactic, semantic and pragmatic values of music semiotics seem to be merged in the message and the messenger, with no regard to the fact, that its significant (pragmatic) value can only be analysed considering the addressee of this message, the listener.

In some writings on music-analysis, some authors link their personal view of a work with the biography of its composer. Charles Fisk's connected the famous left hand thrill in bar seven of Schubert's B flat piano sonata with the composer's supposed homosexuality [8]. This is actually a classic example of dramaturgy of music happening within the listener's mind universe (its own *phaneron*, to use Peirce's jargon [9]). In this case, the listener is Fisk himself. He does not speak about his emotions hereby though, and obviously, his vision of the work can cause major differences in the appreciation of Schubert's sonata in other listeners. He conceives his analysis as a 'story' 'a naively poetic description of what happens in the music' [25]. Fisk concludes his article with the sentence 'What Schubert's last Sonata might hold for me', adding the two last words to the title of the article.

To summarise this section: human reactions are always related to the pragmatic aspect of musical semiotics, that aspect directly linked with the understanding of the link between the musical signs (musical syntactics) and their combinations (musical semantics) in a musical discourse. This is closely related to Peirce's seminal work in the field of semiotics: according to Peirce, signs cannot have a definite meaning,

because meaning ought to be qualified continuously [7] [11]. A musical discourse can be therefore understood only *pragmatically*, that is, after being experienced, and that experience conducts to human reactions of different type, of which emotions is one of the most common, but not the only one.

2.4 Mind Games: Categorisation and Memory Retrieval

Back in 1953, in his work *Philosophische Untersuchungen*, Wittgenstein discussed the matter of categorisation [18][12]. As Levitin explains [6], Wittgenstein took the category **'**game**'** and demonstrated that there is no unequivocal way of describing the word, and that this category could be subscribed to many different items, which all could be recognised as such, but which may not have a direct connection among themselves. This is against the way Aristotle analysed categories. In the Aristotelian thought, *"categories were assumed to be a matter of logic, and objects were either inside or outside a category"* [6]. This means, that they have to be clearly defined, and no fuzzy boundaries among them should exist. Game was Wittgenstein**'**s chosen category to challenge classical categorisation, but that can indeed happen with any other. In other words, what for Aristotle could only be black or white (something belongs or not to a given category), it lost after Wittgenstein its absolute meaning, to turn into a more comprehensive way of dealing with categorising, with the addition of all nuances in-between that the Aristotelian analysis was missing. Wittgenstein proposed that not definition but family resemblance is what characterises category membership [6].

Wittgenstein's approach to categorisation was further developed in the 70s by Rosch [26], with the Prototype Theory, which allows categories to have fuzzy boundaries: objects could be part of many different categories at once, depending on how the object is understood or considered. This theory suggests *"the constructivist view, that an abstract generalization of the stimuli we encounter becomes stored*" [6]. In other words, the abstraction of experience in the form of a prototype or tendency is what it is stored in the brain. This abstraction is contrary to record-keeping memory theories, which say that every single action in our lives is stored in some part of the brain.

Smith et al [27] proposed another view with the Exemplar Theory, based in the storage of specific instances (the 'exemplars' of the name). This theory puts the accent on the residual trace in memory, a record-keeping based theory. The main feature of this theory is that it brings *context* to the discussion: "Under it, details and context are retained in the conceptual memory system" [6]. This is the reason why this theory proposes that new information will be normally evaluated by comparison to existing categories and how closely the new information resembles already known members of the existing category.

From 1997 onwards, research by Nadel [28] (among others), proposed a consolidation model, best known as Multiple-Trace Memory model (MTT), in which, both models seem to converge. MMT explains how the hippocampus is involved in both the storage and retrieval of episodic memory (vital therefore for understanding any kind of dramaturgy), while the neocortex is in charge of semantic

[11] Encyclopaedia Britannica Library - 2004: Peirce [7].

[12] 'Categorization is the process in which ideas and objects are recognized, differentiated and understood'. http://en.wikipedia.org/wiki/Categorization [18].

memory (what has also an impact on how to understand the dramaturgy of events). MTT actually takes elements from both the Prototype and the Exemplar models. Levitin explains, that

> [I]n this kind of models, each experience we have is preserved with high fidelity in our long-term memory system. Memory distortions and confabulations occur when, in the process of retrieving a memory, we either run into interference from other traces that are competing for our attention" …. "or some of the details of the original memory trace have degraded due to normally occurring neurobiological processes". [6]

MTT models indicate that potentially every single memory can be encoded in our memories. And this happens in many parts of the brain, not exclusively in one or two, what would explain why people suffering from amnesia, can remember some aspects of their lives and complete forget about others.

In any case, one of the most interesting issues about MTT models is that they do preserve *context*, that is, not only the exact information of retrieval, but also the context in which it was acquired. This should be vital to the issue of dramaturgy of music and the way a listener categorises what is being listened, to elaborate a story of its own. As seen in section 2.2, the brain is specially fitted to create 'stories'. The left part is mainly the one in charge of that function, and probably the region called orbito-frontal cortex [6]. Therefore, from the point of view of neuroscience, we might say that dramaturgy of music is the story that our brain imagines, a story triggered by the act of listening to music. Just like we instantaneously normally 'invent' a story of someone we just met by reading the facial expressions, so do we too, when we listen to music. And if the human brain does indeed deal with categories at all times –and that is the way we come to understand the world every instant, by ordering our thoughts in different 'files'– this process of categorisation cannot be the exception while listening to music.

The subject of categorisation with regard to music dramaturgy can be linked with the Intention/reception project (IR) by Weale. This project "*situates its primary point of departure in aspects of Landy's research, in particular the issues of access and appreciation in E/A art music. It includes the development, enhancement and expansion of two of his concepts: the 'something to hold on to factor', and 'dramaturgy' in E/A music*" [5]. Even though this definition refers only to electroacoustic music, it can be actually used for any other type of music. I am mostly interested hereby in the concept of '*something to hold on to factor'* and its link with the dramaturgy of music, as it appears to be directly related to categorisation. *"Simply put, the 'something to hold on to factors' (SHFs) are those factors that a listener uses to make sense of and appreciate a particular work"* [5]. Landy made a list of these with different categories [3][5]. In 2005, Weale established a new way of categorising the SHFs [5], enhancing the list proposed by Landy in 1994.

Weale [5] puts the 'dramaturgic information as a SHF'. I would hereby argue that this view seems to imply that the only way to understand the dramaturgy of this music is through the 'dramaturgic information' given by the author of the piece(s) in this research. Whilst in my own categorisation [1], I do not deny that most pieces do have a dramaturgic intention, and that it is vital for a piece of music to be understood as close to its author's conception as possible, I also explain [1], that the emergent dramaturgy in the listener's mind does not need be the one intended at all (and in many cases, it may not coincide at all). However, I totally agree with Landy and Weale in considering the title of the piece a SHF. Their research shows results that seem to prove that the title is a big help in orientating listeners in what they are

about to listen to. However, this does not mean that the title would reveal the entire intention of the piece. And further, it does not mean that this help would always be an aid to find a close understanding of the intended content; it is just a tool of orientation. My own categorisation of emergent music dramaturgy (Fig. 4), situates emotions and thoughts before the emergent dramaturgy in the communication chain, implying that the cultural and emotional baggage of the listener will interact with the input and produce a dramaturgy of its own. MMT, as explained above, seem to support this view.

This said, SHFs can only work, if the brain –while listening to music- react by categorising what is being listened to with previous experiences (regardless of their context). This is directly related to the way information is stored in the brain, as categorisation cannot happen without a known and recognisable background. Memory theories are also linked to categorisation, as we saw above. Levitin explains with quite clearly selected music examples the two main ways of analysis: the constructivist theory (close to Prototype Theory mentioned before), which considers memory as an abstract generalisation of past experiences stored in the brain and not an accurate storage of all of them as the record-keeping theory accounts for (Exemplar Theory) [6]. One of Levitin's examples in favour of the constructivist view, is that people are able to recognize a piece of music in different versions, even transposed to other keys, instrumentation, tempo and variations of its rhythmic (or even form) structure. On the other hand, contextual exemplars do exist while listening and ordering the listened experiences, and are also important. In this way, it is clear that MTT are more flexible models, as they try to incorporate both views, the constructivist (abstract) and the record-keeping.

To explain how the role of categorisation can be linked to understanding music's dramaturgy and human emotions, yet a further, deeper view into the brain's structure is needed. It would appear, that perception and imagination share the same area of the brain. Since the mid 90s and using the help of EEG[13], Janata, doctor in the fields of cognitive neuroscience and neuroethology, studied the relation between imagination and how the brain perceives sound. Janata explains [29]:

> 'Memories of previous sensory input and accumulated knowledge of how the sensory environment behaves are capable of shaping our perceptions of incoming sensory information. Similarly, moment-to-moment sensory input is capable of reshaping stored representations, especially when the recent information doesn't match our expectations'.

Levitin describes an experiment [6], in which he also took part. Janata placed sensors measuring electrical activity from the brain across the surface of the scalp of different test subjects.

> "… Petr and I were surprised to see that it was nearly impossible to tell from the data whether people were listening to or imagining music. The pattern of brain activity was virtually indistinguishable. This suggested that people us the same brain regions for remembering as they do for perceiving". [6]

SHFs may be therefore closely linked to memory issues and could be related to the brain reaction discovered by Janata and Levitin, because the fabrication of stories, as defined above, needs imagination, and listening would appear to share the same part of the brain as imagination. A further categorisation by the brain –whilst listening to music– of a particular SHF by the means of the contextualisation proposed by MTT should follow, linking different categories stored in the memory (and their contexts), to form a particular new story. This process should describe the

[13] Electroencephalography

way we imagine music while listening to it (or even after) and therefore, the process in which music dramaturgy emerges in perception and/or in memory and produces, as a consequence, diverse human reactions. It must be clarified though, that in the *I/R project*, this should apply only to 'reception', not to 'intention'.

4 Conclusion

The semiotics involved in the musical discourse, mostly its pragmatic values, leave an imprint in the human brain and produce what it is called human reactions. These are mainly constituted by thoughts and emotions.

The paper gave a thorough view of the effects those pragmatic values can have on the human brain, by including and explaining the concepts of mental schemas [6], the impact of expectation on them, the different theories about how the human brain categorises (and retains) what it perceives (and considered the MMT model [28] as the most adequate so far to explain those phenomena) and the innate ability of the human brain to imagine stories reflecting its experiences. Emotions and thoughts are therefore included in all of those reactions of the brain to the surrounding world.

With regard to music listening in the field of music dramaturgy via a musical discourse, emotions have been categorised in two ways: either they may predispose the music-recipient to understand the music in a particular way (awake thoughts about what has been experienced emotionally) or they can be aroused by a previous understanding (reflection/thought) of that music. That means, that if the listener is not immediately emotionally involved during the reception of a musical discourse, then thoughts invariably will emerge, as we saw in how the human brain is always prepared to; therefore they are the reaction of the understanding of that particular music. In any of those cases they define (or at least influence) how, during the act of listening, the perception of the music's intrinsic dramaturgy.

Although the two cases exposed in section 2.3 are explained as being completely different, this is so only for the reason of categorisation and clarity. Thus, the most likely situation is that of a rather mixed situation (therefore closer to Wittgenstein rather than to Aristotle), in which the first option may be closer to reality than the second or inversely, the second closer to the first, but never completely and absolutely isolated. As described by Levitin [6] after his experiment with Janata, it is nearly impossible to tell the difference in the data if people were listening to or simply imagining music. The reason given, is that *apparently imagination and listening share the same part of the brain.* Despite the fact, that emotions are simultaneously aroused by other reasons, the main interest of this article relies on the fact that those emotions are linked to that musical imagination.

References

1. Garavaglia, J. A.: Music and technology: What impact does technology have on music's dramaturgy? (invited paper). In: JMM 7, Fall/Winter 2008 - The Journal of Music and Meaning - 2008 (2008), http://www.musicandmeaning.net
2. Garavaglia, J. A.: Music and technology: What impact does technology have on music's dramaturgy? In: Proceedings of the 2008 CMMR/NTSMB Conference - Genesis of

Meaning in Digital Art. K. Jensen (ed.), pp. 99--108. Re:New – Digital arts Forum, Copenhagen (2008)
3. Landy, L.: The 'something to hold on to factor' in timbral composition. In: Contemporary Music Review 10 (2), pp. 49--60. (1994)
4. Landy, L.: Understanding the Art of Sound Organization, MIT Press, London (2007)
5. Weale, R.: Discovering How Accessible Electroacoustic Music Can Be: the Intention/Reception project. In: Organised Sound: Vol. 11, No 2, pp. 189--200. Cambridge: Cambridge University Press (2006)
6. Levitin, D.: This is Your Brain On Music, Atlantic Books, London (2007)
7. Encyclopaedia Britannica Library - 2004
8. Juslin, P., Sloboda, J.: Music and Emotion: Theory and Research, Oxford University Press (2001)
9. The Commens Dictionary of Peirce's Terms. Bergmam, M., Paavola, S. (eds.) (2003), http://www.helsinki.fi/science/commens/terms/phaneron.html
10. Merriam-Webster's Collegiate Dictionary - Encyclopaedia Britannica Deluxe Edition 2004.
11. Misch, I.: Von der Vergangenheit zur Gegenwart. Geschichtsbewusstsein im Schaffen Luciano Berios. In: Musik-Konzepte 128: Luciano Berio. Ulrich Tadday (ed.) Muenchen. Edition text + kritik, pp. 5--21 (2005)
12. Menezes, F.: Das 'laborinthische' Verhältnis von Text und Musik bei Berio. In: Musik-Konzepte 128: Luciano Berio. Ulrich Tadday (ed.) Muenchen. Edition text + kritik, pp. 23--41 (2005)
13. Bergson, H.: Oeuvres. Essai sur les données immédiates de la conscience. Matière et mémoire. Le rire. L'évolution créatrice. L'énergie spirituelle. Les deux sources de la morale et de la religion. La pensée et le mouvant. Presses Universitaires de France, Paris, pp. 783-784. (1970)
14. Pasler, J.: Debussy, "Jeux": Playing with Time and Form. In: 19th-Century Music, Vol. 6, No. 1 (Summer, 1982). University of California Press, pp. 60--75 (1982)
15. Stanford Encyclopedia of Philosophy, http://plato.stanford.edu/entries/events
16. The Internet Encyclopedia of Philosophy, http://www.iep.utm.edu/events/#H1
17. Caesar, M.: Umberto Eco: Philosophy, Semiotics and the Work of Fiction, Blackwell Publishers, Polity Press, p. 81 (1999)
18. http://en.wikipedia.org/wiki/
19. Nattiez, J.: Music and discourse. Toward a semiology of music. New Jersey, Princeton University Press, pp. 127-128 (1990)
20. Kivy, P.: Sound and Semblance. Reflections on Musical Representation. London, Cornell University Press, pp. 3--5 (1991)
21. Scherer, K. R.: What are emotions? And how can they be measured? In: Social Science Information, 44, 4. London, Sage Publications, pp. 695--729 (2005)
22. Delalande, F.: Music Analysis and Reception Behaviours: *Someil* by Pierre Henry. In: Journal of New Music Research 27 (1-2), Routledge, London, pp. 13-66 (1998)
23. Muller, P.: The Story of Read Aloud Virginia. In: Virginia Libraries Journal, Vol. 46, No 3, p. 24. (2000), http://scholar.lib.vt.edu/ejournals/VALib/v46_n3/v46_n3.pdf
24. Budd, M.: Music and the emotions. The philosophical theories. Routledge, London (1994)
25. Fisk, C.: What Schubert's Sonata might hold. In: Music and meaning. J. Robinson (ed.), Ithaca, NY, Cornell University Press, pp. 179--200 (1997)
26. Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., and Boyes-Braem, P. Basic objects in natural categories. In: Cognitive Psychology 8, pp. 382--439 (1976)
27. Smith, E., Medin, D. L.: The exemplar view. In: Foundations of Cognitive Psychology: Core Readings, D. J. Levitin (ed.), Cambridge, MIT Press, pp. 27--292 (2002)
28. Nadel, L. and Moscovitch, M.: Memory consolidation, retrograde amnesia and the hippocampal complex. In: Curr Opin Neurobiol, 7(2), pp. 21--227 (1997)
29. Janata, P.: Electrophysiological Studies of Auditory Contexts. In: The Newsletter, University of Oregon, Eugene, Sept. 1996, Vol. 9, 1, pp. 6-7 (1996) https://scholarsbank.uoregon.edu/xmlui/bitstream/handle/1794/977/96Newsltr.pdf?sequence=1

ENP-Regex - a Regular Expression Matcher Prototype for the Expressive Notation Package

Mika Kuuskankare*

Sibelius Academy, Finland
mkuuskan@siba.fi

Abstract. In this paper we introduce ENP-regex, a prototype of a regular expression matcher developed for Expressive Notation Package (ENP). ENP-regex allows us to use the regular expression syntax to match against several score attributes, such as pitch and rhythm. Instead of writing the regular expression matcher from scratch we implement a scheme where a thin conversion layer is inserted between an existing Lisp-based regular expression library and ENP. The information sent from ENP to the regex matcher is transformed into a textual format. Similarly, the matches are converted into corresponding score objects. The benefit of the present implementation is that potentially the whole syntax of the regex matcher in question is at our disposal. We have implemented a prototype of the regular expression matcher. In this paper we present the current state of the system through examples.

Keywords: Regular expressions, music notation, scripting, music analysis and visualization

1 Introduction

In this paper we present an extension to Expressive Notation Package (ENP, [8]) called ENP-regex. ENP-regex allows us to use regular expressions to match against musical data, such as pitch and rhythm. Traditionally, regular expressions are used for matching characters, words, or patterns of characters in strings. Similarly, with ENP-regex, we are able to match notes, groups of notes and different patterns in an ENP score according to a given property.

Regular expressions and other string search algorithms have been widely used in the music domain. Dovey [4] reports a regular expression like search framework that uses piano-roll notations as a starting point. One of the most notable music analysis applications, Humdrum [6], uses the regular expressions extensively. The problems with representing musical attributes with text are widely discussed in [3].

Our main motivation is to study the potential of regular expressions in the context of ENP. We use symbolic music notation, not text, as a starting point

* The work of Mika Kuuskankare has been supported by the Academy of Finland (SA137619). We would also like to thank CCRMA, Stanford University, for hosting the research.

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

and use musical conventions, rather than textual, when describing the regular expression patterns. This should make the system more approachable for musicians. The mapping between the musical attributes and regular expressions is done on the fly without any further actions required from the user.

A new scripting language is envisioned where any Lisp function could be applied to the matching objects. Potentially, we could insert expressions, add or delete notes, transpose them, etc. For example, an intelligent find (or find and replace) extension could be implemented with the help of regular expressions.

One of the benefits of using regular expressions is that they are widely known and used. The plan is to eventually integrate ENP-regex more closely into the ENP tool-chain.

The rest of the paper is organized as follows. First, we discuss some of the implementation issues. Next, we give some examples of real-world problems where ENP-regex would prove to be useful. The paper ends with some discussion and a list of plans for further development.

2 ENP-regex

ENP-regex is based on a library called cl-ppcre [1] which is a regular expression library for Common Lisp. The ENP-regex matcher can be run in different domains, currently pitch, rhythm, interval, and harmony (pitch-class set), to match against several score properties. The user inputs the regular expression using a slightly modified syntax (this will be discussed below in more detail). The target score is encoded so that it can be processed by the cl-ppcre matcher. The results returned by cl-ppcre (indices) are translated back to score objects, and, finally, the action indicated by the user is performed. At this stage we mark the matches in several different ways, such as inserting expressions, or simply by highlighting the matches.

One of the key concepts behind the ENP-regex implementation is the idea of a translator. A translator maps the desired score objects into a representation that, in turn, can be used as an input to a conventional regex parser, which, in turn, returns indices which are mapped back to score objects. Figure 1 illustrates this process.

The regular expression syntax used in the case of ENP-regex is compatible with that of Perl but slightly extended. Although it would be convenient in our case, normally, we do not write a pattern as $\boxed{\texttt{[60-66]}}$ to match all numbers between 60 and 66. Therefore, for convenience, a small language extension is provided which allows us to use a more musically oriented syntax when defining the regular expression patterns.

For pitch, both absolute pitch and intervals, we use the MIDI note representation, i.e., middle-C is represented by the number 60, and for rhythm the fractional notation, e.g., $\boxed{1/4}$ or $\boxed{1/20}$. Note that our MIDI note representation is extended as it allows us to represent micro-intervals by adding a fractional part, such as 0.5, for example, to denote a quarter tone. For harmony, we use

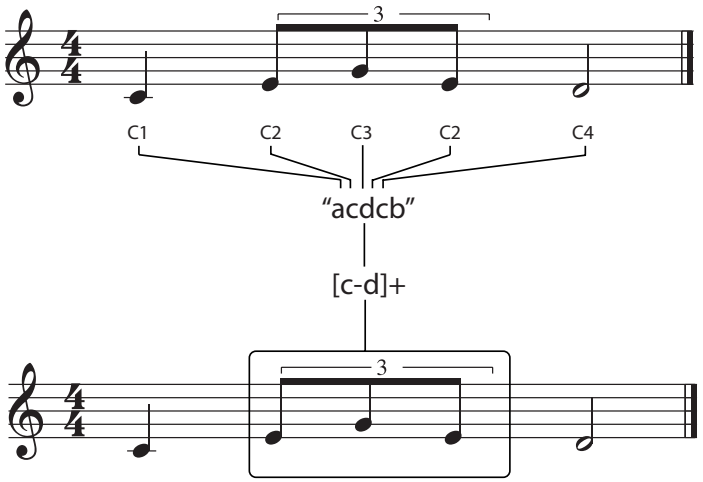


Fig. 1. The translation of score properties into a representation that can be parsed by a regular expression engine and back to score objects.

the pitch-class set notation following the conventions introduced by Allen Forte [2], where the major triad, for example, is notated with the symbol "3-11b" and the minor one with "3-11a".

To distinguish the ENP-regex notation from that of regular expressions we use a hash-mark (#) as a prefix. A pre-processor is implemented which translates our customized regex syntax into the Perl compatible syntax. Thus, it is possible to indicate, for example, a pitch range by writing it as `[#60-#72]`.

3 Examples

In this section we illustrate the potential of ENP-regex through examples.

Figure 2 shows our editor developed for testing the ENP-regex interface. The first row gives the selectors for the property domains, i.e, pitch, rhythm, intervals, and harmony. Using the next group of controllers we can select the desired side-effect. "default" indicates that we want to highlight the matches and "custom" together with the following text input field allow us to specify the class name and the attributes of an ENP-expression [7], which will be applied to the matches found in the score. The third row allows us to choose the matching direction (this will be discussed in Section **??**). Finally, in the bottom row the ENP-regex pattern is given.

3.1 Phrases

We begin with a simple example that aims to illustrate the relationship between regular expressions and ENP. The internal encoding of pitch (and other information) is arranged so that the alphanumeric characters are reserved for attributes

that are related to events, and the 'non-word characters' are reserved for rests. Currently, we do not distinguish between rests of different lengths, but treat them as non-sounding events. Therefore, the 'alphanumeric characters' symbol, `\w`, in ENP-regex is used to indicate a note, and the `\W`, in turn, indicates a rest. We provide this translation for convenience only and it does not attempt to draw any further conclusions about the relationship of music and text.

In our first example (see Figure 2) we use ENP-regex to insert phrasing slurs in the score, using the following regular expression: `\w+`. This is a straightforward way of segmenting music according to the rests. We also use a custom phrasing slur with some additional attributes (see, "SLUR :KIND :DASHED" in Figure 2) instead of simply revealing the matches. The phrasing slur is displayed in the score as a curve using a stippled pattern.

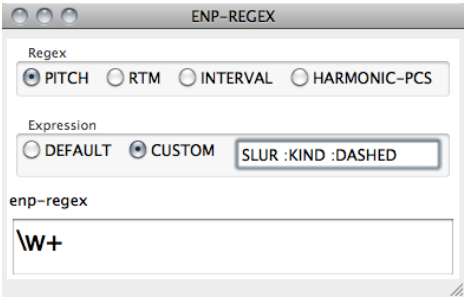


Fig. 2. The ENP-regex tool with the regex expression at the bottom.

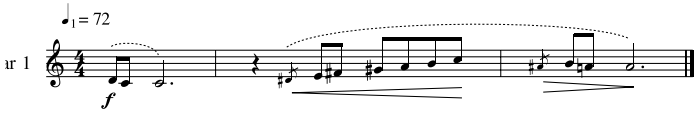


Fig. 3. Inserting phrase marking (the two dashed slurs above the score) with the help of ENP-regex using the pattern \w+. (Yesterday by The Beatles)

3.2 Pitch

In Figure 4, we give an example of ENP-regex in the pitch domain, where we aim to reveal the extreme pitches in a passage written for the flute. The flute spans form B3 to C7 and above. Here, the range considered as extreme is chosen somewhat arbitrarily. The ENP-regex pattern to find and mark the ranges is as follows: [#59-#60#90-#96].

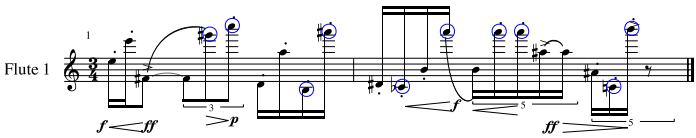


Fig. 4. Indicating extremely low and high pitches (the encircled notes) in the piece of music for the flute using a pattern with low and high ranges: [#59-#60#90-#96].

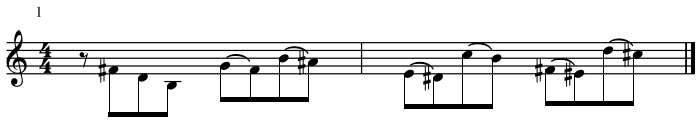


Fig. 5. Articulation slurs inserted according to the interval between two consecutive notes. (J.S.Bach)

3.3 Intervals

As an example of ENP-regex in the interval domain we attempt to add appropriate articulation slurs to a small excerpt of music by J.S. Bach (see Figure 5). We note that in the original there is a slur between two notes forming a descending minor second interval. We define the interval pattern as `#-1` and define the slur expression as in Figure 3 but without the extra attributes (:kind :dashed). Figure 5 shows the slurs inserted with the help of ENP-regex.

3.4 Harmony

The example shown in Figure 7 is a small excerpt, prepared by the Finnish composer Kimmo Kuitunen, called "6-Z47B"-blues. Here, we use ENP-regex in the harmony domain to locate certain sonorities, namely the "mother" set-class 6-Z47B and the set-class named 5-35 (a chord consisting of only perfect fifth intervals is possible to construct using the set-class 5-35). The ENP-regex is given in Figure 6. This example demonstrates the ENP-regex can also be executed in non-metrical context.

3.5 Repetitions Using Back-references

Our final example in this section deals with repetitions. Here, we use a regular expression construct, called a back-reference, which is defined as follows: `(.+)\1+`. The matching is done in the harmony domain. In Figure 8 the matching harmonies are indicated by enclosing them inside boxes. Note, that harmony here is a harmony class, thus it does not have anything to do with a particular setting or voicing. This simple pattern finds repeating series harmonies. The first of the matches is an alternating pattern between two different harmonies, and the latter two matches represent static repeating harmonies.

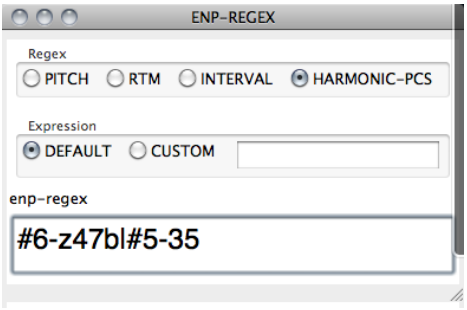


Fig. 6. ENP-regex in the harmony domain aimed at finding and marking specific harmonies in the target score.

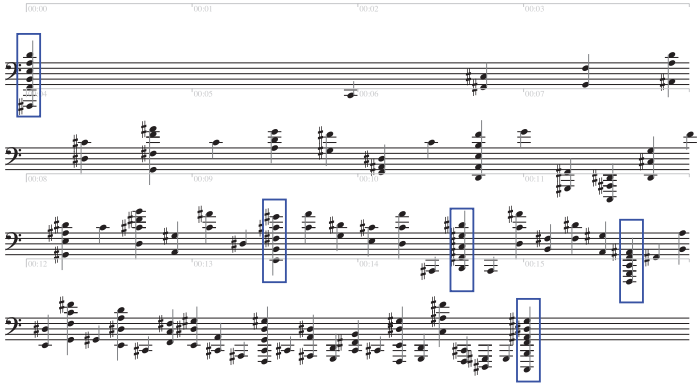


Fig. 7. The harmonies 6-Z47B and 5-35 marked in the score by Kuitunen.

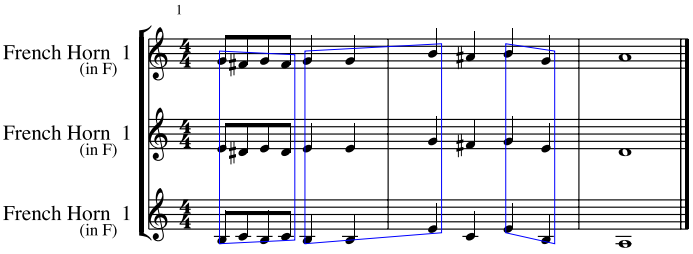


Fig. 8. Harmonic repetitions revealed with the help of ENP-regex using back-references. (Prokofiev: Peter and the Wolf)

4 Future Development

There are several improvements in the planning. First, we should develop a user API for creating custom mappings to any score property, e.g., ENP-expressions.

Second, we should support iterating over higher-level objects than notes. The user could be presented with a choice between notes, chords, and measures, for example. This way we would be free of adding any complexity in the regex pattern, in terms of dealing with the beat boundaries, for example.

Third, we should augment the regular expression specification of ENP-regex. Several additions are planned, such as specifying the matching direction, e.g., right-to-left or bi-directional matching, and incorporating loop-like constructs, such as the beginning index of the matching, step, etc. The latter would allow us to use regex matching to insert, for example, interval n-grams into the target score. n-grams have been have been widely used in text retrieval and are also proposed for MIR applications, for example, in [5].

Finally, the regex matchers could potentially also be combined. It should be investigated if there is a feasible way to logically combine the results. A simple intersection might not be enough, as, for example, a regex `[60-67]{4}` would not necessarily be true, when an intersection is taken with the results returned by the regex `1/4+` executed in the rhythm domain, etc. However, it would be interesting to provide users with the choice as it would allow us to make multi-parameter regular expression matching.

Finally, our regex implementation could also potentially be coupled with the existing pattern matching language of PWGLConstraints[9], thus allowing us to use both syntaxes interchangeably. Some patterns would be more easily expressed using the regex syntax, rather than that of our backtracking constraints system.

5 Conclusions

This paper presents ENP-regex, the prototype implementation of a regular expressions matcher for the Expressive Notation Package. Currently, ENP-regex is able to use most of the regular expression syntax and can match against different types of score information, such as pitch, rhythm, intervals, and harmony.

The most interesting applications of the present work can be found in the domains of music information retrieval, scripting, and computer assisted composition and analysis.

References

1. ppcre. http://weitz.de/cl-ppcre/
2. Allen Forte: The Structure of Atonal Music. Journal of Music Theory (1973)
3. Cambouropoulos, E., Crawford, T., Iliopoulos, C.S.: Pattern processing in melodic sequences: Challenges, caveats and prospects. In: In Proceedings of the AISB'99 Convention (Arti Intelligence and Simulation of Behaviour. pp. 42–47 (1999)

4. Dovey, M.J.: A technique for regular expression style searching in polyphonic music. In: International Symposium on Music Information Retrieval (2001)
5. Downie, J.S.: Evaluating a Simple Approach to Music Information Retrieval: Conceiving Melodic N-grams as Text. Ph.D. thesis, University of Western Ontario (1999)
6. Huron, D.: Music information processing using the humdrum toolkit: Concepts, examples, and lessons. Computer Music Journal 26(2), 15–30 (2002)
7. Kuuskankare, M., Laurson, M.: ENP-Expressions, Score-BPF as a Case Study. In: Proceedings of International Computer Music Conference. pp. 103–106. Singapore (2003)
8. Kuuskankare, M., Laurson, M.: Expressive Notation Package. Computer Music Journal 30(4), 67–79 (2006)
9. Laurson, M.: PATCHWORK: A Visual Programming Language and some Musical Applications. Studia musica no.6, doctoral dissertation, Sibelius Academy, Helsinki (1996)

The Role of Musical Features in the Perception of Initial Emotion

David Taylor[1], Emery Schubert[1], Sam Ferguson[2], Gary McPherson[3]

[1] Empirical Musicology Group, University of New South Wales, Sydney, Australia
[2] University of Technology, Sydney, Australia
[3] University of Melbourne, Melbourne, Australia
david.anthony.taylor@gmail.com

Abstract. 170 participants were played short excerpts of orchestral music and instructed to move a mouse cursor as quickly as possible to one of six faces that best corresponded to the emotion they thought the music expressed. Excerpts were analysed and the musical cues coded. Relationships between the number of cues and participants' response times were investigated and reported. No relationship between the number of cues available to the listener and the speed of response was found. Findings suggest that the initial response to ecologically plausible musical excerpts is quite complex, and requires further investigation to provide emotion-retrieval models of music with psychologically driven data.

Keywords: Music, emotion, modelling, initial response, cue utilisation, response speed

1 Introduction

Modelling emotional responses to music has developed considerably in the last ten years [1]. We now have models that can predict a typical emotional response to a piece of music expected from an individual in a Western culture reasonably accurately, simply by processing certain features extracted from the same piece [2]. Furthermore, use of continuous response methods has allowed understanding and modelling of moment-to-moment changes in musical features and subsequent emotional response [3-6]. Of the many dilemmas these approaches have left, however, continuous responses have highlighted a curious problem regarding the initial response to a piece of music. Recent research suggests that it takes some eight seconds after the start of a piece of music before a listeners emotional response 'settles' or becomes reliable [7, 8].

Since emotional responses appear not to be immediate according to continuous response studies, but can nevertheless be performed quite quickly according to post-performance response data, the question of how quickly one can decide on emotion expressed by music becomes an inviting and relevant question.

Peretz et al. [9] reported that participants were able to differentiate correctly between 'happy' and 'sad' emotions based on mode and tempo as quickly as 0.25 of a second. In a similar study, Bigand et al. [10] compared participants' grouping of one-second excerpts into groups of similar emotional character with the grouping of

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

twenty-five-second excerpts. Strong correlations between the groupings for each excerpt duration confirmed the Peretz et al. [9] findings that only a very small amount of time is needed to induce strong emotional responses. Although these studies demonstrate the extreme rapidity with which emotional responses can be made, the area of reaction time, or 'response' time is one that has remained relatively unstudied and calls for further insight. Bachorik et al. [7] looked at the response time (or what the authors refer to as 'integration time') for participants whom were expressly instructed to 'move [a] joystick as soon as they began feeling an emotional response to the music' whilst their movements were plotted on a two-dimensional grid of arousal and valence. The 'integration time' was measured as the time taken for participants to make a movement of the mouse beyond a pre-determined 'jitter' of 15 pixels. However the study only reports typical integration times (between 8.31s and 11s) and again fails to scrutinise individual results that are faster than this, or indeed, faster than the one-second reported in Bigand et al. [10].

On further scrutiny of the musical content (or 'musical and psychoacoustical structures') of their excerpts, Bigand et al. [10] suggested that the responses were governed by highly cultural compositional and performance-related features or cues. They discovered that many of the one-second excerpts (half of all cases) contained only a single chord or interval, and some only a single pitch and concluded after a 'cautious analysis' that performance cues within the music are enough to induce emotions in Western listeners at this very quick speed. Similar studies that yield comparable response times in making non-emotional evaluations of music would tend to support this [11]. By examining both performance-rated and non-performance-related cues from the position of the fastest possible emotional responses it may be possible to build a clear picture of a) which cues are utilised b) the number necessary in order to make a decision and c) how cues are utilised (i.e. based on their 'usefulness' in any hierarchical structures). By looking at fastest plausible response times it is possible to say with some degree of confidence 'this much music was required before the participant was able to make an emotional response'. By then carefully analysing the musical content and coding the cues with it, it may be possible to start to see more clearly what type or number of cue listeners most rely on.

Substantial work has been undertaken to identify the factors within musical structures that allow us to perceive emotional expression. For example, fast tempo has been associated with the expression of emotions such as 'exciting', 'happy' and 'glad' whereas slow tempo has been associated emotions such as 'serenity' and 'sadness' (for a summary of musical features from reviewed studies see: [12]). Juslin asserts that the number of cues impinges directly on the music's effectiveness to communicate emotion [13], however little research has attempted to reconcile these findings with the initial moments of an emotional response in the listener. This paper will, therefore, ask a number of questions: How quickly is it feasibly possible to recognise emotions in music? Does the number of musical features or 'cues' have any effect on the response time? How many cues are needed before a listener can make an informed decision? The overarching hypothesis is that there exists a cue accumulation effect whereby ambiguous cues (i.e. cues that provide information that conflicts with that from the majority of others) delay the evaluation of cues already made available by requiring further cues in order to confirm an assessment resulting in longer response times.

2 Method

2.1 Participants

A total of 170 participants took part in the experiment, 101 female and 69 male. All were tertiary level students either undergraduate or postgraduate between the ages of 17 to 50 (median age = 21) with a mean average of 6.99 years musical experience (range = 0 – 30 years).

2.2 Materials

The experiment used a pool of 19 short excerpts (duration = 7 s – 27 s) taken from the soundtracks of Disney Pixar animated films. It was deemed that music from this genre had high potential in best conveying the target emotions due to their programmatic and narrative nature. A pilot study was conducted to ascertain which emotion each excerpt was deemed to express the best. In this pilot study, participants were asked to select from a list of emotions (excited, happy, calm, sad, scared or angry) the one they thought was most applicable to each excerpt used in the current trial. The results were consolidated to arrive at a putative emotion for each excerpt (this became the 'target emotion'. See: [14]). All excerpts were purely instrumental. Each target emotion had 3 excerpts from which the software (written by author SF using Max 5) used in the trial could select, with the exception of the excited target emotion, which contained an extra excerpt in order to avoid better the possibility of participants guessing the last target emotion to be played. The software would randomly select one excerpt at a time from each of the target emotions whilst displaying a question at the top of the screen asking participants to identify the emotion they think the music best expresses (as opposed to the emotion the music makes them feel).

2.3 Procedure

Participants were presented with an on-screen display of six 'target faces' arranged in a circle. Each face was a simple representation of a target emotion (either excited, happy, calm, sad, scared or angry) and ordered according to valence and arousal (level of valence arranged horizontally and arousal vertically so that target emotions of similar valence and arousal are adjacent). To further aid quick reference and elicit the fastest possible response, the faces were also coloured (red for angry; yellow for excited, happy and calm; blue for sad, and darker blue for scared). Participants were explicitly instructed to move the mouse cursor to their decision as quickly as possible. Through headphones, they were then played a series of seven excerpts randomly drawn by the software from the pool. The participant initiated playback of each excerpt by clicking a green quaver symbol in the centre of the circle. The computer

then logged each participant's time for their 'out of box' time: the time it took them to move the cursor out of a hidden centre box, and their 'first face' time: the time it took them to move the cursor to the face they first selected. The software also logged all faces 'visited' and in which order together with all other cursor movements. It is to these tasks and responses that the present study refers.

3 Results and Discussion

Data were separated into two groups: data from those participants whose first face was that of the target emotion, plus or minus one (one face either side of the target face); and data from those whose first face was not (but instead was in another location on the screen, including one of the three remaining faces). The second group was eliminated from further analysis. The reason for this was because we are interested, in this study, in the fastest plausible response time after the music begins.

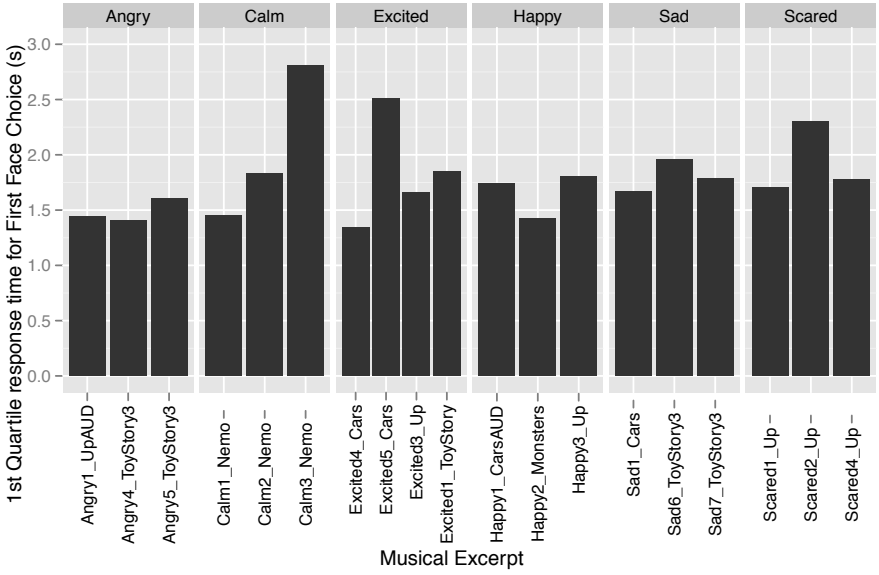


Fig. 1. Comparison of first quartile 'first face' times between individual excerpts

Fig. 1 shows the response times for each individual excerpt by target emotion (the emotion we supposed that the participant would select). First quartile response times (the first face time of the participant who was 25% of participants behind the fastest participant's first face time) were examined because we were interested in the fastest plausible time that participants could make emotional responses. Median time, for example, gives an estimate of typical response time, and minimum response time is

susceptible to error, for example due to prevarication, random error or guessing. First quartile time was therefore deemed a conservative and reasonable comprise (what we term 'plausibly fastest'). From a null-hypothesis perspective, each response time within a target emotion would be identical, which seemed to be the case for 'angry' excerpts and for 'sad' excerpts. Yet some excerpts in other emotions, such as excerpt 3 for 'calm' and excerpt 2 for 'excited' and 'scared' stand out.

Therefore, the question arose, considering that both calm excerpts 1 and 3 were putatively considered to be good examples of music that expressed that emotion in the pilot study, why was there such a difference in response times? The difference suggested that something in the music of the first excerpt made it easier to judge quickly the intended expression. These two excerpts were therefore selected for closer analysis.

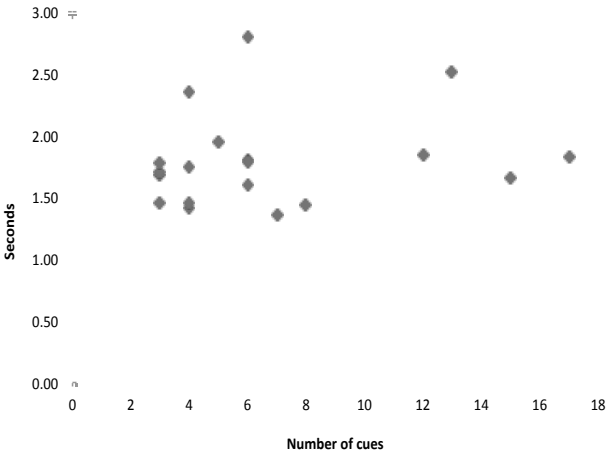


Fig. 2. Scatter chart depicting relationship between 'first face' first quartile response times for each excerpt and number of cues available to the listener (total number of cues counted in that excerpt before the first quartile time)

We were interested to see whether the number of cues available to the listener was in any way related to the speed at which responses were made, i.e. the more cues there were available to the participant, the quicker their response. The first level of cue analysis involved a very basic level of coding. Using 'Audacity', each excerpt was analysed in spectrographic form. A time label was entered at each occurrence of an 'event' (e.g. a note, chord, cymbal clash) between the start of the excerpt and first quartile time. At this level of coding, vertical events (i.e. simultaneous notes from different instruments or chords) were treated as one event. It was not possible to situate cues that unfolded in time in one location, i.e. with regard to a change in loudness where the actual cue started or ended, or indeed where the information from the cue was gleaned. Therefore, for simplicity, cues of variance were omitted at this

stage. .txt files of the cue labels and times were then exported to facilitate counting. Once the cues for each excerpt had been counted they were plotted on a graph along with the first quartile time (Fig. 2).

The main hypothesis was that there would be a correlation between the number of non-ambiguous cues available to the participants and the speed of their response. However, Fig. 2 shows that, at least with a basic count of cues, this did not appear to be the case. If it were, we might expect a trend to be visible where as the number of cues increases, the response time decreases. The first quartile time for Angry1_Up was one of the fastest (at 1.45s), yet participants only received three cues before that time, whereas the fastest 25% of participants responding to Excited5_Cars required thirteen unambiguous cues before feeling able to make a decision (managing only a first quartile time of 2.51s).

This assumes however, that the spacing of cues was consistent and regular, but it may be that in some cases, for example, very few cues were available in the first second followed by a many number in quick succession. Also, clearly the slower the response time, the more cues there will be counted – simply due to the duration. If Angry1_Up had, for example all three cues in the first second, and Excited5_Cars had ten of its thirteen in the first second yet still yielded the slow first quartile time of 2.51s, it would be much stronger evidence against the number of cues being important. Therefore, we counted the number of cues available in each excerpt within the first second only.

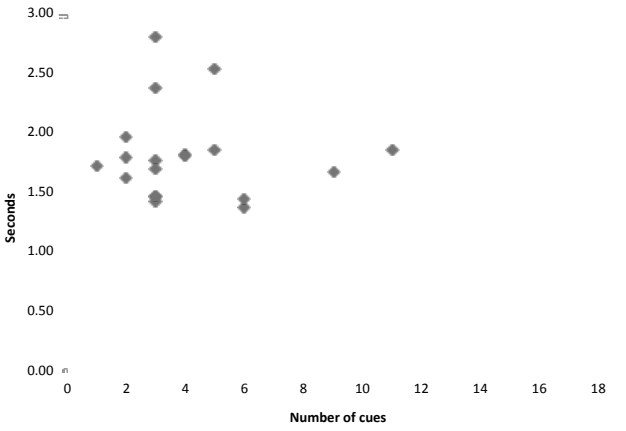


Fig. 3. Scatter chart depicting relationship between 'first face' first quartile response times for each excerpt and number of cues available to participants within the first second

Again, there was no link between the number of cues available to the participant in the first second and the response times. Furthermore, Fig. 3 shows that six of the excerpts all contained three cues in the first second of music but yielded response times ranging from 1.40s to 2.79s.

4 Conclusions and Further Research

The present study examined timing and number of cues and suggests that the initial emotional response to a piece of music is quite complex, and requires considerable further research. Further research is required to examine whether the quality of the cues are relevant in determining response speed. For example, some cues might have a greater weighting than others, giving rise to a hierarchical cue structure, as is the case in some theories of music cognition [15]. While other studies described above have identified rapid identification of emotions in the order of one second, our study was conducted with ecologically typical musical extracts, allowing the underlying complexity of emotion response time to the start of a piece of music to emerge. This has important implications for automated modelling of emotion in music systems because until we can retrieve the underlying nature of human emotional response to music at this transitional, initial orienting period, it will be difficult for time varying models to produce psychologically plausible representations of emotions at the start of a piece of music.

Acknowledgments. This research was funded by the Australian Research Council (DP1094998).

References

1 Schubert, E.: Continuous self-report methods. In: Juslin, P.N., & J. A. Sloboda (eds.) Handbook of music and emotion, pp. 223-253. Oxford University Press (2010),
2 Hug, A., Bello, J.P., and Rowe, R.: Automated Music Emotion Recognition: A Systematic Evaluation. Journal of New Music Research, 39, (3), pp. 227-224 (2010)
3 Schubert, E.: Modeling percieved emotion with continuous musical features. Music Perception, 21, pp. 561-585 (2004)
4 Schubert, E.: Continuous measurement of self-report emotional response to music. In: Juslin, P.N., & Sloboda, J. A. (eds.) Music and Emotion: Theory and research, pp. 393-414. Oxford University Press (2001)
5 Schubert, E.: Measuring emotion continuously: Validity and reliability of the two-dimensioned emotion-space. Australian Journal of Psychology, 51, pp. 154-165 (1999)
6 Schubert, E.: Measurement and time series analysis of emotion in music. Unpublished doctoral dissertation, University of New South Wales (1999)
7 Bachorik, J.P., Bangert, M., Loui, P., Larke, K., Berger, J., Rowe, R., and Schlaug, G.: Emotion in Motion: Investigating the Time-Course of Emotional Judgments of Musical Stimuli. Music Perception, 26, (4), pp. 355-364 (2009)
8 Schubert, E.: Reliability issues regarding the beginning, middle and end of continuous emotion ratings to music. Psychology of Music, In Press
9 Peretz, I., Gagnon, L., and Bouchard, B.: Music and emotion: Perceptual determinants, immediacy, and isolation after brain damage. Cognition, 68, pp. 111-141 (1998)
10 Bigand, E., Vieillard, S., Madurell, F., Marozeau, J., and Dacquet, A.: Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts. Cognition & Emotion, 19, pp. 1113-1139 (2005)
11 Gjerdingen, R.O., and Perrot, D.: Scanning the dial: The rapid recognition of music genres. Journal of New Music Research, 37, (2), pp. 93-100 (2008)

12 Gabrielsson, A., and Lindstrom, E.: The role of structure in the musical expression of emotions. In: Juslin, P.N., & Sloboda, J. A. (eds.) Handbook of Music and Emotion, pp. 367-400. Oxford University Press (2010)
13 Juslin, P.N.: A Brunswikian approach to emotional communication in music performance. In: Hammond, K.R., and Stewart, T.R. (eds.) The Essential Brunswik: Beginnings, Explications, Applications, pp. 426-430. Oxford University Press (2001)
14 Schubert, E., Ferguson, S., Farrar, N., and McPherson, G.E.: Sonification of Emotion 1: Film Music. In: 17th International Conference on Auditory Display (ICAD 2011). Budapest, Hungary: International Community for Auditory Display (ICAD) (2011)
15 Lerdahl, F., and Jackendoff, R.: A Generative theory of tonal music. MIT Press (1983)

Sonic Choreography for Surround Sound Environments

Tommaso Perego

Goldsmiths University of London

mup01tp@cgold.ac.uk

Abstract. A practice-based project that explores design theories of spatial music with choreographic concepts and practices. Aspects of Wave Field Synthesis and Ambisonics technologies are discussed. A production of original works for dance and music in surround sound constitutes the expected outcomes.

Keywords: movement, dance, expression, surround sound, Ambisonics, Wave Field Synthesis, choreography, perspective

1 Introduction

This paper is a presentation of the concepts and investigative ideas that constitute the foundation of an ongoing practice based PhD research entitled 'Sonic Choreography for Surround Sound Environments'. The project explores the direct comparison of artificial sound movement generated by surround sound technologies to dance movement theories and practices. Through collaborative work with choreographers, it will favour the formulation of sonic choreographic concepts in the context of a music composition.

2 Methods

2.1 Aesthetic Research Practice Based

The whole project is an aesthetic research, it focuses on music composition and in particular on the applicability of movement qualities to sound. The aim of the research is to challenge existing practices by exploring them in full, to individuate how musical scopes could be related to sonic choreographic effects in a clear and satisfying manner. The collaboration with choreographers and dancers should work as immediate visual contrast to sound spatial design, which, if it is consistent enough, would counteract the presence of the performers in the space and the ideas of the choreographers.

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

2.2 Collaborative Work

The Game of Life Wave Field Synthesis (WFS) system consisting of 192 speakers [1] will be used in several sessions, to create a piece for dance and music by June 2013. As part of the project a session of three days has already taken place, and more are planned throughout the year.
At Goldsmiths University Digital Studios in London [2] two works with Ambisonics technology are planned: the first work is based on 1st Order Ambisonic [3], to be tested at the studios and in binaural, and the second for Higher Order Ambisonics (HOA) [3] for an horizontal speakers array (for which setup is yet to be found or arranged). Other institutions are considered.
The interest in developing two works for Ambisonics is due to the differences in image resolution between 1st Order Ambisonics and HOA which shall be discussed in this paper.

2.3 Interviews

This research involves contribution from many people including engineers, composers, scientists, scholars and choreographers. Carrying out interviews to gain further knowledge from those people with relevant expertise will add depth to the research and thus it is a vital part of the project. This method is in its arrangement phase.

3 Movement

A general concept of movement can be very difficult to define in many disciplines.
What type of phenomenon is movement and why does it hold such importance in the arts?

3.1 In Dance

There is an extensive literature of studies on movement, most of them written by dance practitioners. During the 1920s, Modern Dance flourished in Germany [4], and later in America and the rest of Europe. The common thread in this pioneering age is that movement speaks to the audience.
With Laban's words [5]: "Movements can be executed with differing degrees of inner participation and with greater or lesser intensity. They may be accelerated by an exaggerated desire to reach a goal or retarded by a cautious doubting attitude. The mover may be entirely concentrated on a movement and use the whole body in an act of powerful resistance, or casually employ only part of the body with delicate touch.

Thus we get different dynamic qualities. One of the basic nuances always shows clearly distinguishable mental and emotional attitudes." [6], and more: "Some of the simplest correlations in space and expression can be described and comprehended without any knowledge of fundamental spatial laws. For instance, when a movement is accompanied by a secondary one in another part of the body in an opposite spatial direction, it can easily be understood that the secondary movement might inhibit or disturb the main movement.[...]Sometimes in this way dynamic nuances can be explained by the spatial influence of secondary movements and tensions."[6]. Laban's *Dynamosphere*, *Kinesphere* and *Effort* theories describe life as a stream of movements, that contains and expresses emotions.

A former Laban student, Mary Wigman, a talented artist who pioneered the *Ausdrucktanz*, stated: "Almost everything that is said about space can also be applied to energy, since energy comes from space" [7], and "The absolute dance is independent of any literary-interpretive content: it does not represent, it is; and its effect on the spectator who is invited to experience the dancer's experience is on a mental-motoric level, exciting and moving" [8].

Breaking away from of the "Expressive" dance movements, more recent creators like Cunningham where found describing: "Dancing provides an amplification of energy that is not provided any other way, and that's what interests me" [9], "There is an ecstasy in dance beyond the idea of the movement being expressive of a particular emotion or meaning. There can be an exaltation in the aura that the freedom of a disciplined dancer provides, that is far beyond any literal rendition of meaning" [10].

Through these examples it appears that movement is transmitting or carrying something valuable, and transferring this valuable energy to people in the form of emotion, experience and artistic expression.

3.2 Expression

The meaning of expression is a critical topic, the sole term has been subject of many philosophical and semiotic studies. The sentence "Music is the relation between sound and intellect" [11] contains a deep anthropological insight as it describes what is found as musical as relative to the human intellect. Stravinsky highlighted in his Poetics of Music [12], that music has no expression: "in the pure state music is free speculation", which confirms the relativeness of the concept of expression and its derivation from subjective appreciations/dislike factors out of the artist control.

Research in music psychoanalysis has proven how far we are from having found stable solid notions for analysis, such as a neutral level [13] that would act as a starting point for a useful observation of musical *phenomena*. This is still missing and far from being defined, and methodologies are hence struggling to be successful.

Stravinsky again mentioned the artisan role of the composer [12], the *homo faber*, sculpting sound material, which turns the attention on skill, mastery, seen as the only real tangible thing against the relativeness of expression. The reality of music is the

astonishing combinations a composer can create with sounds, not what the sound means to anybody. Similarly, no matter what the expressive content of a movement could be, the mastery of it constitutes a central assessment that needs to be investigate through practice and observation.

4 Sound Movement

When the concept of movement is applied to sound, several distinctions must be made. For instance we are referring to the movement of sound within a surround sound environment, then for movement of sound it is intended the movement of a virtual source, not a movement embedded into the sound (e.g. a sound like in a recording of a park, the sudden movement of a bicycle), neither the sound wave acoustical motion. Hence it could be helpful to refer to it as artificial sound movement.
Source bonding concepts [14] refer to sound as a carrier of meaning, which is inseparable and determines our understanding of it. Issues about direction, proximity and individuality are well described in [15], and affect the way we create and experience a music composition. The way in which sound image is represented into the diffusion space is subject of accurate studies, whilst sound movement in current literature appears to be relegated in a secondary position, like a less important part of the overall sound image. Sometimes virtual sources movement practices are referred to as successful, but unsatisfying the complex and full dimension of sound reality [16]. Yet this is understood as confined in personal artistic interests, or sound material choices, which doesn't truly contradict the artistic relevance of movement of sound.
This research wants certainly to assess clarity and accuracy of the sonic image, and at the same time to observe what attracts our attention in movement, in whichever form it becomes manifest (as meaning/expression/feeling/energy/ecstasy...), and how it relates to the music discourse.
When a sound image appears in space, it is choreographically relevant: when a movement burst out, it dictates symmetry, correspondence as it is engaging the space. The connection with its content, a relation of consonance and amplification or of reduction, contradiction, exaggeration of the sound reality and material, is part of the artistic game, is how we want it to be or appear. When this connection comes to life, its a tangible sign of the value of the existence/presence of the movement in the scene.

4.1 Decoding

The first issue is about how sound is diffused in the room. Different technologies have different approaches for rendering the spatial attributes of sound. In this research I take in consideration Wave Field Synthesis and Ambisonics, and in particular: 1st

order Ambisonic in 3D, Higher Order Ambisonics (realistically for an horizontal array only), WFS synthesis of 192 speakers setup in a 10x10 meter space.
A loudspeakers system should be designed to give a realistic, natural impression of space [16], and how it delivers this is crucial for the appreciation of movement as it is for the sound image.
All of these techniques are not flawless, and many situations affect the clear perception of sound movement, too many to be included in this paper. There is a general assumption that, if a sound image is clear, choreography possibilities are then absolute and unlimited. That is doubtful because of the presence of several perceptual artefacts in movement process itself, and moreover because the system design imposes a dominant sound projection perspective [20].
What it is here relevant to note is thus that the general rendition of sonic movement through different decoding systems may have different results, which directly affect the choreographic potential.

Ambisonics. Ambisonics technology works in different resolutions, and this affects the way we experience sound. In 1st Order Ambisonics, the listening area is very limited, emphasising the importance of a sweet spot. In HOA this tends to be reduced because many other spherical harmonics are added to increase resolution, so that they cover a wider listening area. Yet rendition of sound outside the sweet spot still introduces artefacts, that affects the scene intelligibility, and generally it is thought that the best perceptual location is to stand in the middle of the speakers perimeter [15, 17 and 19]. Because of this, questions arise on how to show dance in the listening space, and how sonic choreography could adapt to a dance in this space, given its ties with Ambisonics spherical sound projection. It appears to me there is a perspective conflict that as to be taken in consideration while composing, to be accommodated artistically. Whilst music content could be infinite and borderless, sonic choreography is more likely to be limited by and constructed around system characteristics. According to the space and technology available a specific approach should be favoured. This project is thus trying to build a sonic choreography on the 1st Order Ambisonics characteristic emphasising a centric perspective of sound projection through artistic exploration; a second experiment would be with HOA technology, that will require as well a specific strategy. The point is that a sonic choreography seems limited in its design as it responds to a perspective, and this perspective is dictated by the system used, like in a dancer the body is the limit and range of the expression.

Wavefield Synthesis. “Wave field synthesis (WFS) is a spatial sound field reproduction technique that utilizes a high number of loudspeakers to create a virtual auditory scene over a large listening area. It overcomes some of the limitations of stereophonic reproduction techniques, like e. g. the sweet-spot” [18].

This clearly puts WFS in a different context in respect to Ambisonics, and comparisons are difficult and maybe pointless: this other paper [17] it's indicative of the irreducible differences between the two systems, Ambisonics and WFS, in terms of math, limits of sound image rendition (artefacts differences), and quality of representation of the sound field.
For a dance performance with WFS, the accompanying sonic choreography is surrounding the audience, as listeners have to be inscribed into a perimeter of speakers, but no center is needed for the perception, which allows several dispositions of dancers, sounds and choreographies, including a frontal display of choreography, which resemble the more common way to experience a dance performance. Whilst WFS seems not to affect the choreographic design by an imposed perspective of sound projection, how the impression of depth is rendered has yet to be explored in the project, especially how different locations and perspectives and relation between locations of sound sources are actually perceived.
On these considerations it is in development a work plan for the collaboration with choreographers in this particular space, for which the aim is to significantly differentiate from the Ambisonics approach.

4.2 Encoding

For creating movements of sound, so called spatialisation tools within an authoring framework [20] are needed. Movement design tools, for drawing trajectories of sound in space should make available any type of geometrical operation, grouping/singling of sources and trajectories, the possibility of a single and multiple virtual source representation of the same sound, tools for dealing with speed, including correction or realism of doppler effects and the relation with amplitude, gain attenuation and air absorption filters, synthesis of early reflections and reverberation: all these processes should be easily accessible by a composer, when inventing sonic choreographies; all those sonic issues should also be addressed, for a consistent representation of sound movement. For example, the current availability of plugins for encoding provides a standard three coordinates system or azimuth and elevation, which satisfies mainly the generic positioning of sounds in space. Although it is possible to use them to design trajectory of movement, more sophisticated design tools are needed.
The project Holo Edit [21 and 20] seriously provided a model, based on a Digital Audio Workstation (DAW) structure, for unify under the same interface a set of geometrical and spatial transformations and a communication system through OSC [22] protocol. Others platforms and plugins are worth mentioning (Open Music, the ICST tools for Max/MSP, Jamoma, WigWare, Harpex-b and Ambisonic Studio), but still they don't represent a unified approach that satisfies movement design to its core. This research, through direct comparison with dance, is trying to highlight these problems as well to find solutions for them.

The SpatDIF project is proposing a SDIF format for storing spatial information [20] and OSC for streaming data, which is embedded within HoloEdit and few others softwares. This should overcome the limit imposed by system design to aesthetic invention as mentioned also here [20, 2.2] .
Those two projects are based on a team of artists and researchers: that is encouraging as for the production of surround sound the development of movement practises and technology should not go without the composers input and experience, and I'd add as well of choreographers and motion experts.

5 Conclusion

This research is posing technical and theoretical interrogatives for the analysis of a creative process involving music composition for surround sound environments. The artistic meaning of sound movement is investigated, for which the realistic possibilities are assessed for the different systems used. The hope is, through interviews and demonstrations and with an aesthetic and compositional approach, to contribute in reaching the core of the problems to the improvement of the operational framework, knowledge and artistic potential of sonic movement design.

References

1. The Game Of Life, http://gameoflife.nl
2. Digital Studios, Goldsmiths University Of London, http://www.gold.ac.uk/gds/
3. Ambisonics General Theory references, http://www.york.ac.uk/inst/mustech/3d_audio/gerzonrf.htm, http://en.wikipedia.org/wiki/Ambisonics
4. Anderson, Jack: Art without boundaries. London: Dance Book, Cecil Court (1997); Howe, Dianne S.: Individuality and Expression: the Aesthetics of the New German Dance, 1908-1936. New York: Peter Lang Publishing (2001); Toepfer, Karl: The Empire of Ecstasy: Nudity and Movement in German Body Culture, 1910-1935 (1997)
5. Laban, R - Ullmann L. : Choreutics. London: Macdonald and Evans (1966)
6. Laban, R - Ullmann L. : Choreutics. London: Macdonald and Evans, 27 (1966)
7. Wigman Mary: Vom Studium des modernen Tanzes. Berlin: Notebook, MWA
8. Reynolds, Dee. : Rhythmic Subjects. Uses of energy in the dances of Mary Wigman, Martha Graham and Merce Cunningham, Alton: Dance Books Ltd, 61 (2007)
9. Cunningham, cited by Susan Sontag, in Conversation on the dance, Cunningham, Merce - Sontag, Susan, recorded 3rd March 1986 (audiocassette, Dance Collection at New York Public Library for the Performing Arts at Lincoln Center; henceforth NYPL)
10. 'Excerpts from lecture-demonstration given at Anna Halprin's Dance Deck (13 July 1957), in David Vaughan, Merce Cunningham, 101 (100-01).
11. Celibidache, Sergiu in Schmidt-Garre, Jan: Celibidache:You don't do anything, you let it evolve. Music Documentary. ParsMedia and ZD5 [Video:DVD] (1992)

12. Stravinsky, Igor: Poetics of Music: in the form of six lessons. Cambridge, Massachussets and London, England: Harvard University Press, 47-64 (2003)
13. Delli Pizzi, Fulvio: SEMEION/TECMERION: Verso una psicanalisi dlla musica. Milano: Libreria Clup, Segni e Suoni - SIMC, 12-18 (2007)
14. Smalley Denis: Space-form and the acousmatic image. Organised Sound: Vol. 12, No. 1. Cambridge: Cambridge University Press: 35-58 (2007)
15. Barrett, Natasha: Ambisonics spatialisation and spatial ontology in an acousmatic context. Proceedings from the Electroacoustic Music Studios conference. <http://www.natashabarrett.org/EMS_Barrett2010.pdf>, 4-5 (2010)
16. Gerzon, Michael: Wither four Channels? Audio Annual 1971, Croydon: Link House Publications, 36-41 (1971)
17. Spors, Sascha and Ahrens, Jens: A Comparison of Wave Field Synthesis and Higher-Order Ambisonics with Respect to Physical Properties and Spatial Sampling. Proceedings of 125th AES Convention (2008)
18. Spors, Rabenstein, Ahrens: The Theory of Wave Field Synthesis revisited. Proceedings of 124th AES Convention, 1 (2008)
19. Bates, Enda: The Composition and Performance of Spatial Music. Ph.D. Thesis, Trinity College, Dublin (2009)
20. Peters, Lossius, Schacher, Baltazar, Bascou, Place: A stratified Approach, For Sound Spatialisation. Proceedings of The 6th Sound and Music Computing Conference, 23-25 July 2009, Porto.< A Stratified Approach For Sound Spatialization> (2009)
21. Bascou, Charles: Adaptive Spatialisation and Scripting Capabilities in the Spatial Trajectory Editor Holo-Edit. Proceedings of SMC Conference 2010 <http://smcnetwork.org/files/proceedings/2010/59.pdf> (2010)
22. Open Sound Control < http://opensoundcontrol.org/introduction-osc>

An Investigation of Music Genres and Their Perceived Expression Based on Melodic and Rhythmic Motifs

Debora C. Correa[1], F. J. Perez-Reche[2] and Luciano da F. Costa[1]

[1] Instituto de Fisica de Sao Carlos, Universidade de Sao Paulo, Sao Carlos, SP, Brazil
[2] SIMBIOS Centre, University of Abertay, Dundee, UK
deboracorrea@ursa.ifsc.usp.br, p.perezreche@abertay.ac.uk, luciano@ifsc.usp.br

Abstract. The constant growth of online music dataset and applications has required advances in MIR Research. Music genres and annotated mood have received much attention in the last decades as descriptors of content-based systems. However, their inherent relationship is rarely explored in the literature. Here, we investigate whether or not the presence of tonal and rhythmic motifs in the melody can be used for establishing a relationship between genres and subjective aspects such as the mood,dynamism and emotion. Our approach uses symbolic representation of music and is applied to eight different genres.

Keywords: music genres, melodic motifs, rhythm, mood.

1 Introduction

Online music data has significantly increased in number and size in the last decade. Specially, web radios and online stations have received much attention, due to recent research involving music recommendations systems applied for large-scale music collections. In this scenery, music genres together with mood in music are particularly interesting descriptors, since they summarize common characteristics of music and are included in the set of principal tools for content-based music retrieval and organization.

There are many previous works dealing with the task of automatic classification of music genres [1]. There is also some work related to the classification of mood in music [2]. Mood and emotion classifications are challenging tasks, since they involve subjective notions and face the difficulty of establishing an accepted taxonomy of mood and emotions.

The inherent correlation of music genres and mood have not been very much explored in the literature. In [3], the authors obtained substantial improvement in music emotion classification by including the genre information audio songs. The work of Hu and Downie [4] also explores the relationships of mood-genre, mood-artist, and mood-recommendation usages. They applied statistical analysis to metadata collections like *All Music Guide.com*, *epinions.com* and *Last.fm*,

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

demonstrating that important evidences in genre-mood and artist-mood relationships could be used in the development of a more succinct dataset of "mood-spaces" that minimizes redundant problems in emotional terms.

Within this context, the paper aims at contributing to the existing investigation of how music genres can be related with mood and thus establish complementary descriptors that can be used to improve current applications of music information retrieval systems. Our method is applied to MIDI files and based on derived temporal configuration patterns in the melodies of songs, also known as *motifs*. We analyse the presence of tonal and rhythmic motifs, demonstrating that they can be linked to the way we describe or feel a specific genre. Our motivation for analyzing the melodies is associated with the fact that the melody is one of the first music aspects that make us recognize a song. In addition, it is the contour of a melody what we usually first memorize from a song [5].

The remainder of the paper is organized as follows. Section 2 describes the method and the used dataset. Section 3 dwells the principal results and discussion. Finally, Section 4 presents the concluding remarks.

2 Methods

The proposed method is summarized in Figure 1 and detailed in the following.

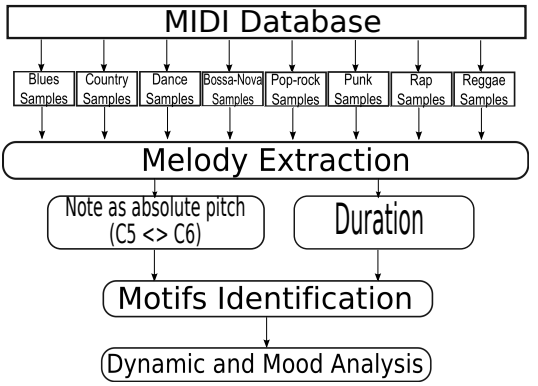


Fig. 1. The proposed method. After selecting MIDI songs from different genres, the voice related to the melody is extracted and represented by a vector of absolute pitches and by a vector of the note values. Absolute pitches are indicated as MIDI numbers. For example, C5 is 72 and C4 is 60. Note values are represented as relative durations (e.g. the eighth note takes the value 0.5 and the quarter note takes the value 1).

2.1 Data Description

Our database consists of eight music genres, namely, blues (34 songs), country (30 songs), dance (29 songs), bossa-nova (Brazilian music, 29 songs), punk (40

songs), pop/rock (39 songs), rap (5 songs), and reggae (12 songs). These genres are widely known and relatively easy to obtain as MIDI files in the Internet with a reasonable quality. We chose to use symbolic representation because it is a compact representation. Specially, the MIDI format offers the possibility of separating the melody voice, providing a deep analysis of the music elements.

To edit the MIDI files, the Sibelius software was used together with the free Midi Toolbox for Matlab computing environment [6]. The voice related to the melody (or singing voice) was extracted in each song and represented as a note matrix. Each column of the note matrix contains information about quantities such as the relative duration (in beats), MIDI channel, MIDI pitch, or intensity.

We propose to analyse the temporal patterns in the melody following two different procedures: considering the absolute pitch (AP), and the relative note value of the pitches (NV). For the AP representation the pitches are the events, for example, C4, D4, F#4, C5, D5 and so on. For the NV case, each event represents one possible note value, such as half note, quarter note, or eighth note. The relative note value is represented in this matrix through relative numbers (for example, 1 for quarter note, 0.5 for eighth note, 0.25 for sixteenth note and so on). In order to deal with possible fluctuations in tempo, we deactivated an option in Sibelius called "Live Playback". In this way, the note values in the MIDI file preserve their relative proportion (e.g., the eighth note is always 0.5).

To illustrate the idea, Figure 2 shows part of the melody of the song *"From me to you"* (*The Beatles*). The respective AP and NV vectors are also presented.

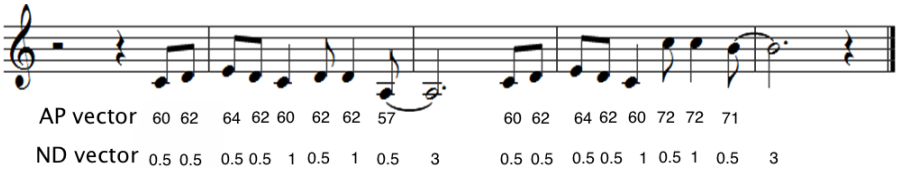


Fig. 2. An example of the representation of the melody of the song "From me to you" (The Beatles) using the AP (absolute pitch) and NV (note value) vectors.

2.2 Finding the Motifs

Music motifs (also known as "motives") are fundamental in music compositions. Basically, there are two forms of constructing music motifs: keeping the tonal sequence and changing the rhythmic structure; or keeping the rhythmic sequence and changing the tonal sequence. We consider both cases in this work. Due to repetitions and returns in popular music, it is also possible to find motifs that retain tonal and rhythmic sequences at the same time.

Our approach to find the tonal and rhythmic motifs in the melodies is relatively straightforward. In order to exemplify the idea, consider the AP vector from the melody in Figure 2. In the first step, we iteratively compare the AP

vector with shifted versions of itself. The comparison is done note by note, as illustrated in Table 1. The size of the motif is determined by the number of notes in the original vector that coincide with those in the shifted vector. For each shift degree, we count how many times motifs of different sizes occur. Table 1 shows that shifting this sequence by a lag of two results in three motifs of size one; while shifting by a lag of eight results in one motifs of size four, representing the main tonal and rhythmic motif of the adopted example.

AP vector	60	62	64	62	60	62	62	57	60	62	64	62	60	72	72	71
shift size 1	-	60	62	64	62	60	62	62	57	60	62	64	62	60	72	72
match	-	F	F	F	F	F	T	F	F	F	F	F	F	F	T	F
AP vector	60	62	64	62	60	62	62	57	60	62	64	62	60	72	72	71
shift size 2	-	-	60	62	64	62	60	62	62	57	60	62	64	62	60	72
matches	-	-	F	T	F	T	F	F	F	F	F	T	F	F	F	F
...																
AP vector	60	62	64	62	60	62	62	57	60	62	64	62	60	72	72	71
shift size 8	-	-	-	-	-	-	-	-	60	62	64	62	60	62	62	57
matches	-	-	-	-	-	-	-	-	T	T	T	T	T	F	F	F

Table 1. Examples of tonal motifs for the melodic sequence in Figure 2.

This process establishes a matrix that we denote as APM with rows and columns representing the shift iterations and the quantity of motifs of different size, respectively. For example, the entry APM(1,3) indicates how many motifs of size three occurred when the AP vector was shifted by a lag of size one. Finally, by calculating the average value of each column in this matrix, we obtain the average frequency that motifs of different sizes occurred in the melody of the corresponding song. Small motifs will not be used in our analysis since they do not necessarily represent a relevant repetitive patterns and are expected to have a highly random character. We arbitrarily consider motifs of size five or higher.

NV duration vector	0.5	0.5	0.5	0.5	1	0.5	1	0.5	3	0.5	0.5	0.5	0.5	1	0.5	1	0.5	3
shift size 1	-	0.5	0.5	0.5	0.5	1	0.5	1	0.5	3	0.5	0.5	0.5	0.5	1	0.5	1	0.5
match	-	T	T	T	F	F	F	F	F	F	T	T	T	F	F	F	F	F
NV duration vector	0.5	0.5	0.5	0.5	1	0.5	1	0.5	3	0.5	0.5	0.5	0.5	1	0.5	1	0.5	3
shift size 2	-	-	0.5	0.5	0.5	0.5	1	0.5	1	0.5	3	0.5	0.5	0.5	0.5	1	0.5	1
matches	-	-	T	T	F	T	T	T	F	T	F	T	T	F	T	T	T	F
...																		
NV duration vector	0.5	0.5	0.5	0.5	1	0.5	1	0.5	3	0.5	0.5	0.5	0.5	1	0.5	1	0.5	3
shift size 9	-	-	-	-	-	-	-	-	-	0.5	0.5	0.5	0.5	1	0.5	1	0.5	3
matches	-	-	-	-	-	-	-	-	-	T	T	T	T	T	T	T	T	T

Table 2. Examples of rhythmic motifs for the melodic sequence in Figure 2.

The rhythmic motifs are obtained following the same idea. Table 2 demonstrates comparisons for the Beatles' melody shown in Figure 2. Comparison of

Table 2 with Table 1 reveal that the information from rhythmic motifs differs from that brought by the analysis of tonal motifs. Shifting the sequence by a lag of nine, it is possible to find the main rhythmic motif of the sequence.

3 Results and Discussion

The presence of motifs was analised for each one of the eight genres. We first investigate whether or not the songs from different genres can be discriminated by the frequency of repeated tonal or rhythmic motifs in their melodies. We then explore the association of motifs with positive or negative emotions based on the link between music factors (rhythm, melody, and musical form) and emotions proposed in [11] (see summary in Table 3).

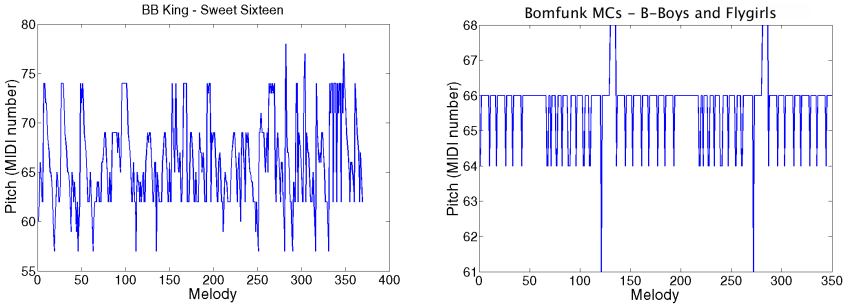
Rhythm	Regular/smooth	happiness, dignity, majesty, peacefulness
	Irregular/complex	amusement, uneasiness, anger
	Flowing/fluent	happy/gay, graceful, dreamy.
Melody	Wide melodic range	joy, uneasiness
	Narrow melodic range	sad, sentimental, delicate
	Stepwise motion	dullness
Musical form	High complexity	tension, sadness
	Low complexity	joy, peace, relaxation

Table 3. Expression of emotions according to different music factors: rhythm, melody and musical form [11].

As described earlier, we calculate the quantity of tonal and rhythmic motifs from different sizes for each genre. The tonal motifs bring information about the tonal contour of the melody. Predictable or modular melodies have constant contours, with many repeated parts. On the other hand, more dynamic contours are usually encountered in melodies with motifs that do not have a high degree of repetition. Figure 3 illustrates the countour of two different melodies, representing the genres blues and rap. While the rap melody is significantly regular, the blues melody is more dynamic and has many variations in the repeated parts.

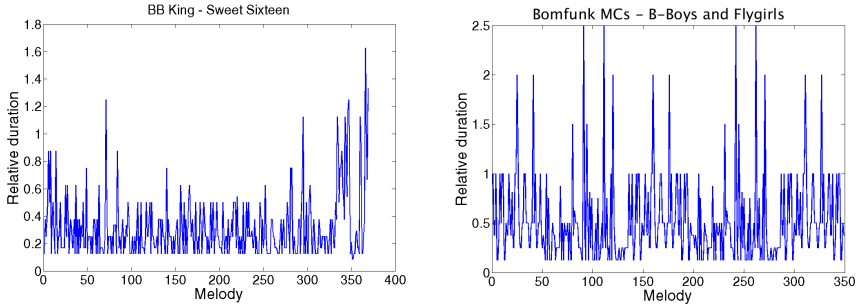
The rhythmic contour of the melodies brings different information. Figure 4 presents the same examples as in Figure 3. The rhythmic structure in blues is significantly dynamic. Rap has a more regular rhythmic pattern, but it is interesting to note that the dynamics of its rhythmic contour differs from its tonal counterpart. Thus, we propose to link both aspects in order to correlate genre and mood. The mood annotations for the genres analysed in this work were mainly obtained from the *All Music Guide* site [7].

Figure 5 (a) and Figure 5 (b) show, respectively, the analysis for the tonal and rhythmic motifs. In both cases, we plotted the motif sizes against the average quantity of times it appears for each genre.



(a) (b)

Fig. 3. The AP vector of two melodies. (a) A blues melody by the music *Sweet Sixteen* by *BB King* and (b) A rap melody by the music *B-Boys and Flygirls* by *Bomfunk MCs*.



(a) (b)

Fig. 4. The NV vector for the melodies in Figure 3.

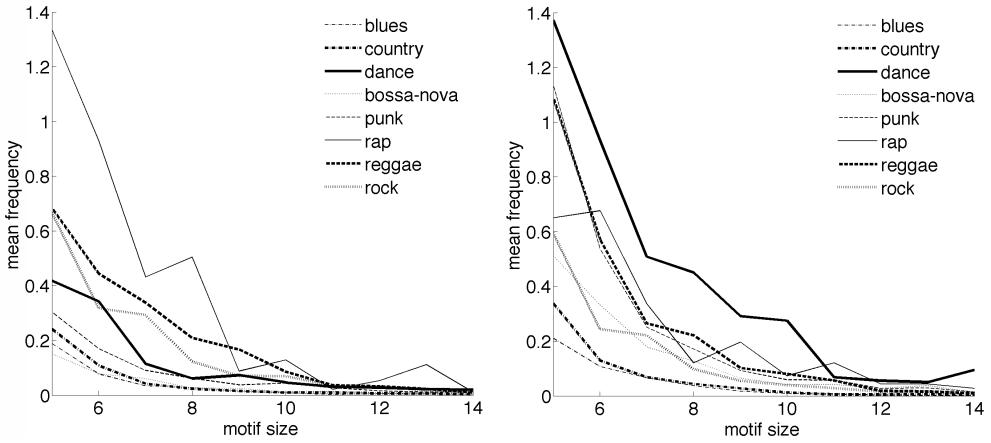


Fig. 5. The configuration of (a) tonal and (b) rhythmic motifs in the melodies of the music genres.

Rap has the higher quantity of tonal motifs, and the second higher quantity of rhythmic motifs. According to [8], rap is known by its chanted rhyming lyrics and regular flow. Rap artists generally receive mood annotations like cheerful, fun, exciting, harsh and angry. When compared to Table 3, the annotations agree with regular rhythm and flowing. Rapping consists of mainly three componentes: content, flow, and delivery [8]. "Flow" is related to the rhythm and rhyme aspects and how they interact, while "delivery" contains elements like pitch, timbre and volume and it is more related to the melody or the form the rap is sung. This may explain why rap has the second highest quantity of rhythmic motifs, differing from tonal motifs, since the tonal motifs mainly represents the spoken characteristic of the genre. The delivery may contain more irregular components, which may contribute to the negative annotations. Rap was influenced by reggae [8]. They share common characteristics like the syncopated and regular rhythms. The melody of reggae is characterised by a simple feel and sense of phrasing [7]. This agrees with the results, since reggae is the second higher in presence of tonal motifs and the third higher in the presence of rhythmic motifs. Common annotations for reggae albums are relaxing, restrained and soothing.

Dance is the most rhythmically repetitive genre in the results. While the tonal sequence may present some variations, dance music has a defined rhythmic beat. This is in agreement with the mood annotations found in [7] such as energetic, happiness and lively. The constancy and modularity in dance music determines its intrinsic attribute: a steady rhythm that stimulates body movements.

Blues and country genres usually do not aggregate many repeated motifs in their melodies, either in tonal or rhythmic aspects. Both genres receive annotations like complex, sophisticated, sentimental and stylish. Country and blues share similar themes and songs, since blues was a stylistic origin of country. The melodies of both genres are characterized by the selection of specific notes (for instance, the flattened third, fifth and seventh) and narrow melodic sequences [10]. This contributes for annotations like sentimental or melancholily. Country songs are formed in simple chords and a plain melody, but these basic forms allow a substantial range of variations and different styles, from resolved patterns to improvisations [7]. This is reflected in the results, since country and blues have a small quantity of repeated motives when compared to the other genres.

Rock music has a defined rhythmic structure but it is usually more dynamic than, for example, dance. Common annotations for rock albums are: energetic, ambitious and exciting. Although referred here as a genre, punk is also known as a rock style, with basic chords progressions and simple melody (played in a louder and faster manner [7]). This is reflected in the results, mainly in the rhythmic analysis of the melody, since punk seems to have a considerable number of rhythmic motifs.

Bossa-nova is a kind of Brazilian music originated from jazz and samba. It is harmonically complex, but its melodies have a constant rhythm [10]. This explains why it is similar to rock when analyzing the rhythmic motifs. However, the lyrics in bossa-nova are usually richer than in rock in terms of tonal variations [10], an aspect well captured by our results regarding the tonal motifs.

4 Concluding Remarks

We proposed a link between music genre and mood using the presence of melodic motifs in the songs. The melody or vocal track is extracted from MIDI files and represented by a vector of note pithes and note values. We derived a method to identify tonal and rhythmic motifs in each melody and relate the frequency of their occurrence to mood notions. For validation purposes, we collected mood annotations from artists in our dataset using the *All Music Guide* site [7].

Genres like rap, dance, reggae and rock, known for their constant rhythmic patterns were found to have a higher quantity of motifs in their melody. Blues and country confirmed their fame to be "more sophisticated" genres, since it is not common to find many motifs that are fully repeated. We expect that such kind of information can help to improve music-content classification systems.

This work represents the first steps of a deeper study which will include a more complete examination of genres and other evaluation methods. In principle, it would be relatively straightforward to include other characteristics of songs in our analysis, such as rhythm of the percussion tracks and instrumentation.

Acknowledgments. Debora C Correa thanks Fapesp financial support under process 2009/50142-0. Luciano da F. Costa thanks CNPq and Fapesp financial support under processes 301303/06-1 and 573583/2008-0, respectively.

References

1. Scaringella, G. Z., Mlynek, D.: Automatic Genre Classification of Music Content: a Survey. IEEE Signal Proc. Magazine 23(2), 133-141 (2006)
2. Huron, D.: Perceptual and Cognitive Applications in Music Information Retrieval. In: 1st International Society for Music Information Retrieval, (2000)
3. Lin, Y-C, Yang, Y-H, Chen, H. H., Liao, I-B, Ho, Y-C: Exploiting Genre for Music Emotion Classification. In: IEEE International Conference on Multimedia and Expo, pp. 618–621, IEEE Press, New York (2009)
4. Hu, X., Downie, J. S.: Exploring Mood Metadata: Relationships With Genre, Artist and Usage Metadata. In: 8th International Society for Music Information Retrieval, pp.67–72, Vienna (2009)
5. Snyder, B.: Memory for Music. In: Hallan, S., Cross, I., Thaut, M. (eds.) The Oxford Handbook of Music Psychology, pp-107–117. Oxford University Press (2009)
6. Eerola, T., Toiviainen, P.: MIDI toolbox: Matlab Tools for Music Research. University of Jyväskylä (2004)
7. All Music Guide, `www.allmusic.com`
8. Edwards, P.: How to Rap: the Art & Science of the Hip-Hop MC. Chicago Review Press, United States (2009)
9. Unterberger, R.: Birth of Rock & Roll, In: Bogdanov, V., Woodstra, C., Erlewine, S. T. (eds) All Music Guide to Rock: the Definitive Guide to Rock, Pop, and Soul, pp-1303–4. Milwaukee (2002)
10. Oxford Music Online, `www.oxfordmusiconline.com`
11. Gabrielsson, A.: The Relationship between Musical Structure and Perceived Expression. In: Hallan, S., Cross, I., Thaut, M. (eds.) The Oxford Handbook of Music Psychology, pp-1041–150. Oxford University Press (2009)

Subjective Emotional Responses to Musical Structure, Expression and Timbre Features: A Synthetic Approach

Sylvain Le Groux[1], Paul F.M.J. Verschure[1,2]

[1]SPECS, Universitat Pompeu Fabra
[2]ICREA, Barcelona
{sylvain.legroux, paul.verschure}@upf.edu

Abstract. Music appears to deeply affect emotional, cerebral and physiological states, and its effect on stress and anxiety has been established using a variety of self-report, physiological, and observational means. Yet, the relationship between specific musical parameters and emotional responses is still not clear. One issue is that precise, replicable and independent control of musical parameters is often difficult to obtain from human performers. However, it is now possible to generate expressive musical material such as pitch, velocity, articulation, tempo, scale, mode, harmony and timbre using synthetic music systems. In this study, we use a synthetic music system called the SMuSe, to generate a set of well-controlled musical stimuli, and analyze the influence of musical structure, performance variations and timbre on emotional responses.The subjective emotional responses we obtained from a group of 13 participants on the scale of valence, arousal and dominance were similar to previous studies that used human-produced musical excerpts. This validates the use of a synthetic music system to evoke and study emotional responses in a controlled manner.

Keywords: music-evoked emotion, synthetic music system

1 Introduction

It is widely acknowledged that music can evoke emotions and synchronized reactions of experiential, expressive and physiological components of emotion have been observed while listening to music [1]. A key question is how musical parameters can be mapped to emotional states of valence, arousal and dominance. In most of the cases, studies on music and emotion are based on the same paradigm: one measures emotional responses while the participant is presented with an excerpt of recorded music. These recordings are often extracted from well-known pieces of the repertoire and interpreted by human performers who follow specific expressive instructions. One drawback of this methodology is that expressive interpretation can vary quite a lot from one performer to another, which compromises the generality of the results. Moreover, it is difficult, even

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

for a professional musician, to accurately modulate one single expressive dimension independently of the others. Many dimensions of the stimuli might not be controlled for. Besides, pre-made recordings do not provide any control over the musical content and structure.

In this paper, we propose to tackle these limitations by using a synthetic composition system called the SMuSe [2,3] to generate stimuli for the experiment. The SMuSe allows to generate synthetic musical pieces and to modulate expressive musical material such as pitch, velocity, articulation, tempo, scale, mode, harmony and timbre. It provides accurate, replicable and independent control over perceptually relevant time-varying dimensions of music.

Emotional responses to music most probably involve different types of mechanisms such as cognitive appraisal, brain stem reflexes, contagion, conditioning, episodic memory, or expectancy [4]. In this study, we focused on the direct relationship between basic perceptual acoustic properties and emotional responses of a reflexive type. As a first approach to assess the participants' emotional responses, we looked at their subjective responses following the well-established three dimensional theory of emotions (valence, arousal and dominance) illustrated by the Self Assessment Manikin (SAM) scale [5,6].

2 Methods

2.1 Stimuli

This experiment investigates the effects of a set of well-defined musical parameters within the three main musical determinants of emotions, namely structure, performance and timbre. In order to obtain a well-parameterized set of stimuli, all the sound samples were synthetically generated. The composition engine SMuSe[1] allowed the modulation of macro-level musical parameters (contributing to structure, expressivity) via a graphical user interface [2,3], while the physically-informed synthesizer PhySynth[2] allowed to control micro-level sound parameters [7] (contributing to timbre). Each parameter was considered at three different levels (Low, Medium, High). All the sound samples[3] were 5 s. long and normalized in amplitude with the Peak Pro[4] audio editing and processing software. .

Musical Structure: To look at the influence of musical structure on emotion, we focused on two simple but fundamental structural parameters namely register (Bass, Tenor and Soprano) and mode (Random, C Minor, C Major). A total of 9 sound samples (3 Register * 3 Mode levels) were generated by SMuSe (Figure 1).

[1] http://goo.gl/Vz1ti
[2] http://goo.gl/zRLuC
[3] http://goo.gl/5iRM0
[4] http://www.bias-inc.com/

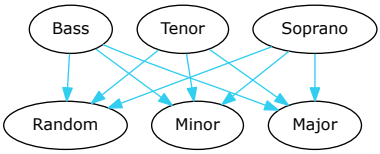


Fig. 1. Musical structure samples: Register and Mode are modulated over 9 sequences (3*3 combinations)

Expressivity Parameters: Our study of the influence of musical performance parameters on emotion relies on three expressive parameters, namely tempo, dynamics, and articulation that are commonly modulated by live musicians during performance. A total of 27 sound samples (3 Tempo * 3 Dynamics * 3 Articulation) were generated by SMuSe (Figure 2).

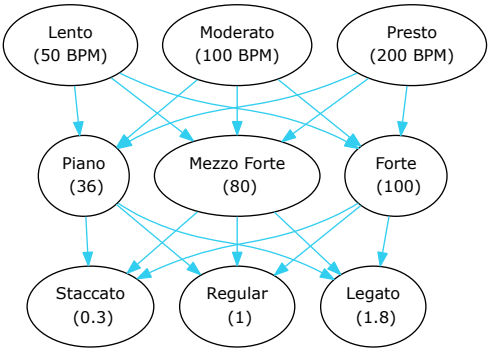


Fig. 2. Musical performance samples: 3 performance parameters were modulated over 27 musical sequences (3*3*3 combinations of Tempo (BPM), Dynamics (MIDI velocity value) and Articulation (duration multiplication factor) levels).

Timbre: For timbre, we focused on parameters that relate to the three main dimension of timbre namely brightness (controlled by tristimulus value), attack-time and spectral flux (controlled by damping). A total of 27 sound samples (3 Attack Time * 3 Brightness * 3 Damping) were generated by PhySynth (Figure 3). For a more detailed description of the timbre parameters, refer to [7].

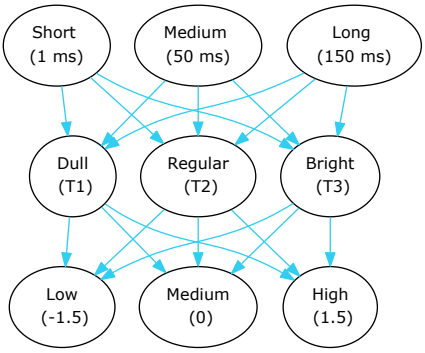


Fig. 3. Timbre samples: 3 timbre parameters are modulated over 27 samples (3*3*3 combinations of Attack (ms), Brightness (tristimulus band), Damping (relative damping α)). The other parameters of PhySynth were fixed: decay=300ms, sustain=900ms, release=500ms and global damping $\alpha_g = 0.23$.

2.2 Procedure

We investigated the influence of different sound features on the emotional state of the patients using a fully automated and computer-based stimulus presentation and response registration system. In our experiment, each subject was seated in front of a PC computer with a 15.4" LCD screen and interacted with custom-made stimulus delivery and data acquisition software called PsyMuse[5] (Figure 4) made with the Max-MSP [6] programming language [8]. Sound stimuli were presented through headphones (K-66 from AKG).

At the beginning of the experiment, the subject was exposed to a sinusoidal sound generator to calibrate the sound level to a comfortable level and was explained how to use PsyMuse's interface (Figure 4). Subsequently, a number of sound samples with specific sonic characteristics were presented together with the different scales (Figure 4) in three experimental blocks (structure, performance, timbre) containing all the sound conditions presented randomly.

For each block, after each sound, the participants rated the sound in terms of its emotional content (valence, arousal, dominance) by clicking on the SAM manikin representing her emotion [6]. The participants were given the possibility to repeat the playback of the samples. The SAM 5 points graphical scale gave a score (from 0 to 4) where 0 corresponds to the most dominated, aroused and positive and 4 to the most dominant, calm and negative (Figure 4). The data was automatically stored into a SQLite[7] database composed of a table for

[5] http://goo.gl/fx0OL
[6] http://cycling74.com/
[7] http://www.sqlite.org/

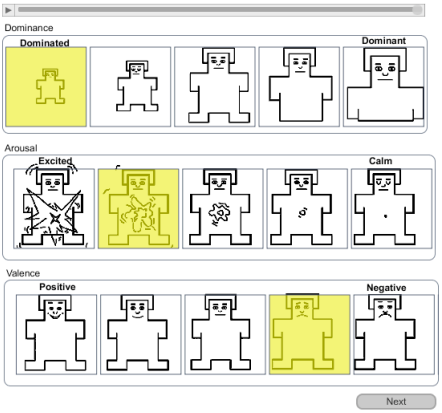


Fig. 4. The presentation software PsyMuse uses the SAM scales (axes of Dominance, Arousal and Valence) [6] to measure the participant's emotional responses to a database of sounds.

demographics and a table containing the emotional ratings. SPSS[8] (from IBM) statistical software suite was used to assess the significance of the influence of sound parameters on the affective responses of the subjects .

2.3 Participants

A total of N=13 university students (5 women, $M_{age} = 25.8$, range=22-31) with normal hearing took part in the pilot experiment. The experiment was conducted in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki[9]. Six of the subjects had musical background ranging from two to seven years of instrumental practice.

3 Results

The experiment followed a blocked within-subject design where for each of the three block (structure, performance, timbre) every participant experienced all the conditions in random order.

3.1 Musical Structure

To study the emotional effect of the structural aspects of music, we looked at two independent factors (register and mode) with three levels each (soprano, bass, tenor and major, minor, random respectively) and three dependent variables (Arousal, Valence, Dominance). The Kolmogorov-Smirnov test showed that the

[8] http://www.spss.com/
[9] http://www.wma.net/en/30publications/10policies/b3/index.html

data is normally distributed. Hence, we carried a Two-Way Repeated Measure Multivariate Analysis of Variance (MANOVA).

The analysis showed a multivariate effect for the $mode * register$ interaction $V(12, 144) = 1.92, p < 0.05$. Mauchly tests indicated that assumption of sphericity is met for the main effects of register and mode as well as for the interaction effect. Hence we did not correct the F-ratios for follow-up univariate analysis.

Follow-up univariate analysis revealed an effect of **register** on **arousal** $F(2, 24) = 2.70, p < 0.05$ and **mode** on **valence** $F(2, 24) = 3.08, p < 0.05$ as well as a **mode** $*$ **register** interaction effect on arousal $F(4, 48) = 2.24, p < 0.05$, dominance $F(4, 48) = 2.64, p < 0.05$ and valence $F(4, 48) = 2.73, p < 0.05$ (Cf. Table 1).

	ANOVAs		
	Register	**Mode**	**Register * Mode**
Arousal	F(2,24)=2.70, *p<.05	NS	F(4,48)=2.238, *p<0.05
Valence	NS	F(2, 24)=3.079, *p<0.05	F(4,48)=2.636, *p<0.05
Dominance	NS	NS	F(4,48)=2.731, *p<0.05

Table 1. Effect of mode and register on the emotional scales of arousal, valence and dominance: statistically significant effects.

A post-hoc pairwise comparison with Bonferroni correction showed a significant mean difference of -0.3 between High and Low register and of -0.18 between High and Medium on the arousal scale (Figure 5 B). High register appeared more arousing than medium and low register.

A pairwise comparison with Bonferroni correction showed a significant mean difference of -0.436 between random and major (Figure 5 A). Random mode was perceived as more negative than major mode.

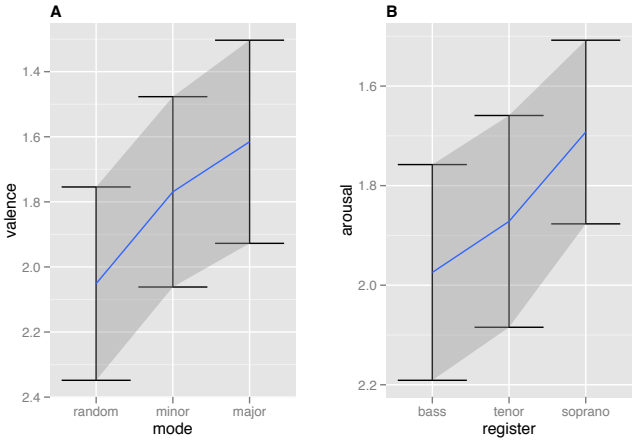


Fig. 5. Influence of structural parameters (register and mode) on arousal and valence. A) A musical sequence played using random notes and using a minor scale is perceived as significantly more negative than a sequence played using a major scale. **B)** A musical sequence played in the soprano range (respectively bass range) is significantly more (respectively less) arousing than the same sequence played in the tenor range. Estimated Marginal Means are obtained by taking the average of the means for a given condition.

The interaction effect between mode and register suggests that the random mode has a tendency to make a melody with medium register less arousing (Figure 6, A). Moreover, the minor mode tended to make high register more positive and low register more negative (Figure 6, B). The combination of high register and random mode created a sensation of dominance (Figure 6, C).

3.2 Expressive Performance Parameters

To study the emotional effect of some expressive aspects of music during performance, we decided to look at three independent factors (Articulation, Tempo, Dynamics) with three levels each (high, low, medium) and three dependent variables (Arousal, Valence, Dominance). The Kolmogorov-Smirnov test showed that the data was normally distributed. We did a Three-Way Repeated Measure Multivariate Analysis of Variance.

The analysis showed a multivariate effect for **Articulation** $V(4.16, 3) < 0.05$, **Tempo** $V(11.6, 3) < 0.01$ and **dynamics** $V(34.9, 3) < 0.01$. No interaction effects were found.

Mauchly tests indicated that the assumption of sphericity was met for the main effects of articulation, tempo and dynamics on arousal and valence but not dominance. Hence we corrected the F-ratios for univariate analysis for dominance with Greenhouse-Geisser.

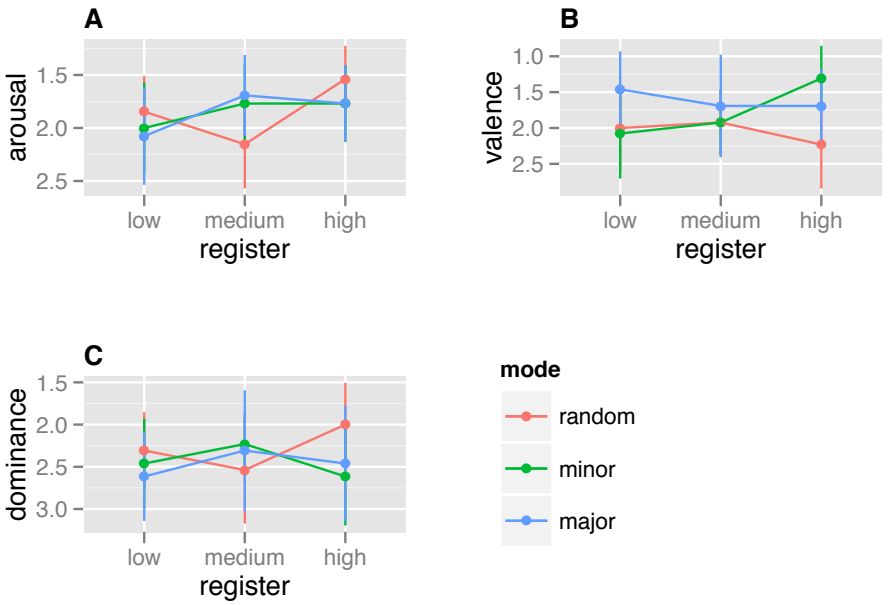


Fig. 6. Structure: interaction between mode and register for arousal, valence and dominance. A) When using a random scale, a sequence in the tenor range (level 3) becomes less arousing **B)** When using a minor scale , a sequence played within the soprano range becomes the most positive. **C)** When using a random scale, bass and soprano sequences are the most dominant whereas tenor becomes the less dominant.

	ANOVAs		
	Articulation	**Tempo**	**Dynamics**
Arousal	F(2,24)=6.77, **p<0.01	F(2,24)=27.1, ***p<0.001	F(2,24)=45.78, ***p<0.001
Valence	F(2,24)=7.32, **p<0.01	F(2, 24)=4.4, *p<0.05	F(2,24)=19, ***p<0.001
Dominance	NS	F(1.29,17.66)=8.08, **p<0.01	F(2,24)=9.7, **p<0.01

Table 2. Effect of articulation, tempo and dynamics on self-reported emotional responses on the scale of valence, arousal and dominance**:** statistically significant effects.

Arousal Follow-up univariate analysis revealed an effect of **articulation** $F(6.76, 2) < 0.01$, **tempo** $F(27.1, 2) < 0.01$, and **dynamics** $F(45.77, 2) < 0.05$ on arousal (Table 2).

A post-hoc pairwise comparison with Bonferroni correction showed a significant mean difference of 0.32 between the **articulation** staccato and legato (Figure 7 A). The musical sequence played staccato was perceived as more arousing.

A pairwise comparison with Bonferroni correction showed a significant mean difference of -1.316 between high **tempo** and low tempo and -0.89 between high and medium tempo (Figure 7 B). This shows that a musical sequence with higher tempi was perceived as more arousing.

A pairwise comparison with Bonferroni correction showed a significant mean difference of -0.8 between forte and piano **dynamics**, -0.385 between forte and regular and 0.41 between piano and regular (Figure 7 C). This shows that a musical sequence played at higher dynamics was perceived as more arousing.

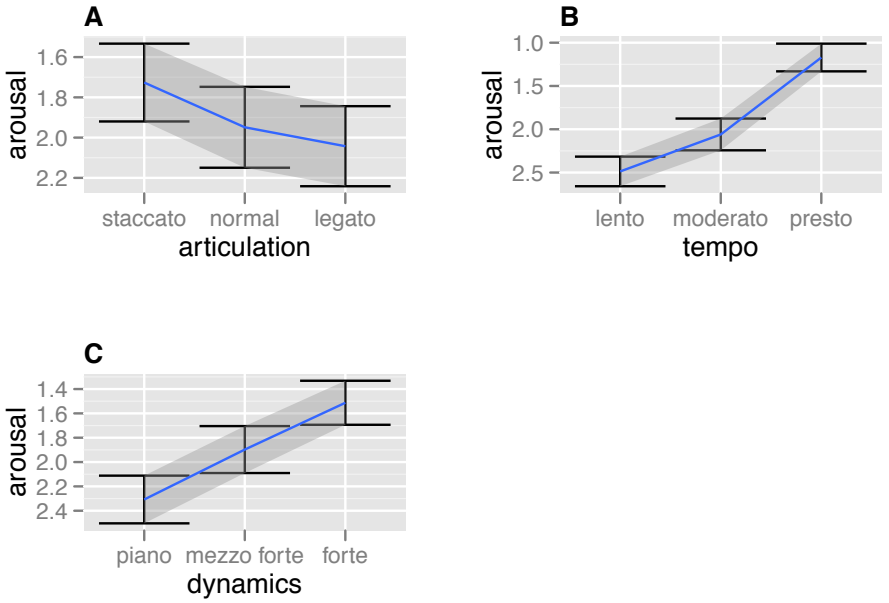


Fig. 7. Effect of performance parameters (Articulation, Tempo and Dynamics) on Arousal. A) A sequence played with articulation staccato is more arousing than legato **B)** A sequence played with the tempo indication presto is more arousing than both moderato and lento. **C)** A sequence played forte (respectively piano) was more arousing (respectively less arousing) than the same sequence played mezzo forte.

Valence Follow-up univariate analysis revealed an effect of **articulation** $F(7.31, 2) < 0.01$, **tempo** $F(4.3, 2) < 0.01$, and **dynamics** $F(18.9, 2) < 0.01$ on valence (Table 2)

A post-hoc pairwise comparison with Bonferroni correction showed a significant mean difference of -0.32 between the **articulation** staccato and legato (Figure 7 A). The musical sequences played with shorter articulations were perceived as more positive.

A pairwise comparison with Bonferroni correction showed a significant mean difference of 0.48 between high **tempo** and medium tempo (Figure 8 B). This shows that sequences with higher tempi tended be perceived as more negatively valenced.

A pairwise comparison with Bonferroni correction showed a significant mean difference of 0.77 between high and low **dynamics** and -0.513 between low and medium. (Figure 8 C). This shows that musical sequences played with higher dynamics were perceived more negatively.

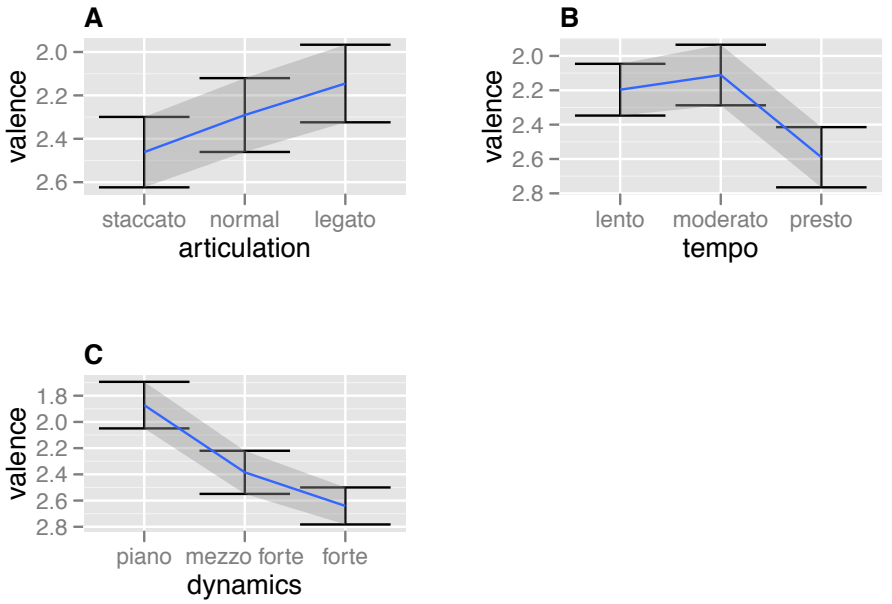


Fig. 8. Effect of performance parameters (Articulation, Tempo and Dynamics) on Valence. A) A musical sequence played staccato induce a more negative reaction than when played legato **B)** A musical sequence played presto is also inducing a more negative response than played moderato. **C)** A musical sequence played forte (respectively piano) is rated as more negative (respectively positive) than a sequence played mezzo forte.

Dominance Follow-up univariate analysis revealed an effect **Tempo** $F(8, 2) < 0.01$, and **dynamics** $F(9.7, 2) < 0.01$ on valence (Table 2).

A pairwise comparison with Bonferroni correction showed a significant mean difference of -0.821 between high **tempo** and low tempo and -0.53 between high tempo and medium tempo (Figure 9 A). This shows that sequences with higher tempi tended to make the listener feel dominated.

A pairwise comparison with Bonferroni correction showed a significant mean difference of -0.55 between high and low **dynamics** and 0.308 between low and medium (Figure 9 B). This shows that when listening to musical sequences played with higher dynamics, the participants felt more dominated.

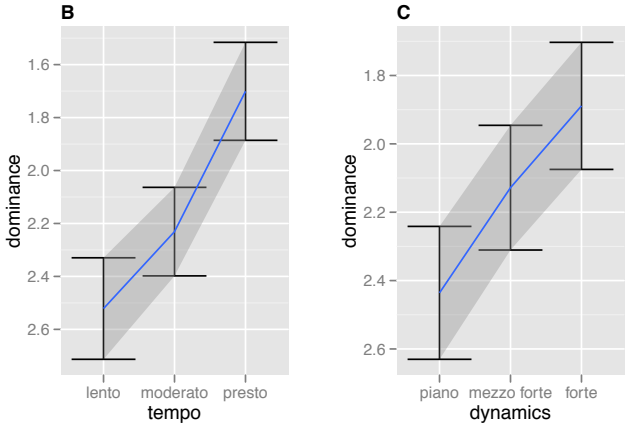


Fig. 9. Effect of performance parameters (Tempo and Dynamics) on Dominance. A) A musical sequence played with a tempo presto (repectively lento) is considered more dominant (respectively less dominant) than played moderato **B)** A musical sequence played forte (respectively piano) is considered more dominant (respectively less dominant) than played mezzo-forte

3.3 Timbre

To study the emotional effect of the timbral aspects of music, we decided to look at three independent factors known to contribute to the perception of Timbre [9,10,11] (Attack time, Damping and Brightness) with three levels each (high, low, medium) and three dependent variables (Arousal, Valence, Dominance). The Kolmogorov-Smirnov test showed that the data is normally distributed. We did a Three-Way Repeated Measure Multivariate Analysis of Variance.

The analysis showed a multivariate effect for **brightness** $V(6, 34) = 3.76, p < 0.01$, **damping** $V(6, 34) = 3.22, p < 0.05$ and **attack time** $V(6, 34) = 4.19, p < 0.01$ and an interaction effect of **brightness** $*$ **damping** $V(12, 108) = 2.8 < 0.01$

Mauchly tests indicated that assumption of sphericity was met for the main effects of articulation, tempo and dynamics on arousal and valence but not dominance. Hence we corrected the F-ratios for univariate analysis for dominance with Greenhouse-Geisser.

	ANOVAs			
	Brightness	**Damping**	**Attack**	**Brightness* Damping**
Arousal	F(2,18)=29.09, ***p<0.001	F(2,18)=16.03, ***p<0.001	F(2,18)=3.54, *p<0.05	F(4,36)=7.47, ***p<0.001
Valence	F(2,18)=5.99, **p<0.01	NS	F(2,18)=7.26, **p<0.01	F(4,36)=5.82, **p<0.01
Dominance	F(1.49,13.45) =6.55, *p<0.05	F(1.05,10.915) =4.7, *p<0.05	NS	NS

Table 3. Effect of brightness, damping and attack on self-reported emotion on the scales of valence, arousal and dominance: statistically significant effects.

Arousal Follow-up univariate analysis revealed the main effects of **Brightness** $F(2, 18) = 29.09 < 0.001$, **Damping** $F(2, 18) = 16.03 < 0.001$, **Attack** $F(2, 18) = 3.54 < 0.05$, and interaction effect **Brightness * Damping** $F(4, 36) = 7.47, p < 0.001$ on Arousal (Figure 3).

A post-hoc pairwise comparison with Bonferroni correction showed a significant mean difference between high, low and medium **brightness**. There was a significant difference of -1.18 between high and low brightness, -0.450 between high and medium and -0.73 between medium and low. The brighter the sounds the more arousing.

Similarly significant mean difference of .780 between high and low **damping** and -0.37 between low and medium damping were found. The more damped, the less arousing.

For the **attack time** parameter, a significant mean difference of -0.11 was found between short and medium attack. Shorter attack time were found more arousing.

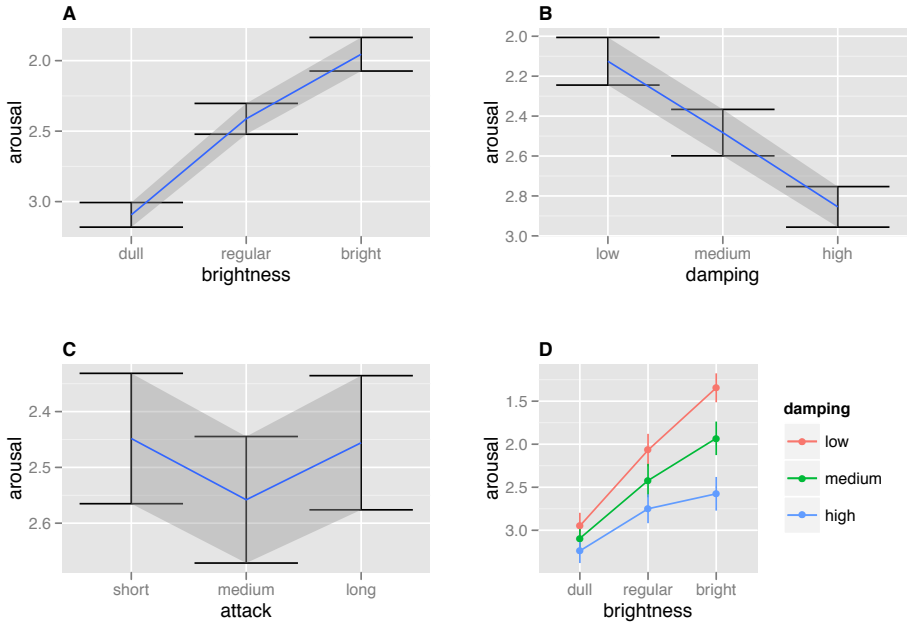


Fig. 10. Effect of timbre parameters (Brightness, Damping and Attack time) on Arousal. A) Brighter sounds induced more arousing responses. **B)** Sounds with more damping were less arousing. **C)** Sounds with short attack time were more arousing than medium attack time. **D)** Interaction effects show that less damping and more brightness lead to more arousal.

Valence Follow-up univariate analysis revealed main effects of **Brightness** $F(2, 18) = 5.99 < 0.01$ and **Attack** $F(2, 18) = 7.26 < 0.01$, and interaction effect **Brightness * Damping** $F(4, 36) = 5.82, p < 0.01$ on Valence (Figure 3).

Follow up pairwise comparisons with Bonferroni correction showed significant mean differences of 0.78 between high and low **brightness** and 0.19 between short and long **attacks** and long and medium attacks. Longer attacks and brighter sounds were perceived as more negative (Figure 11).

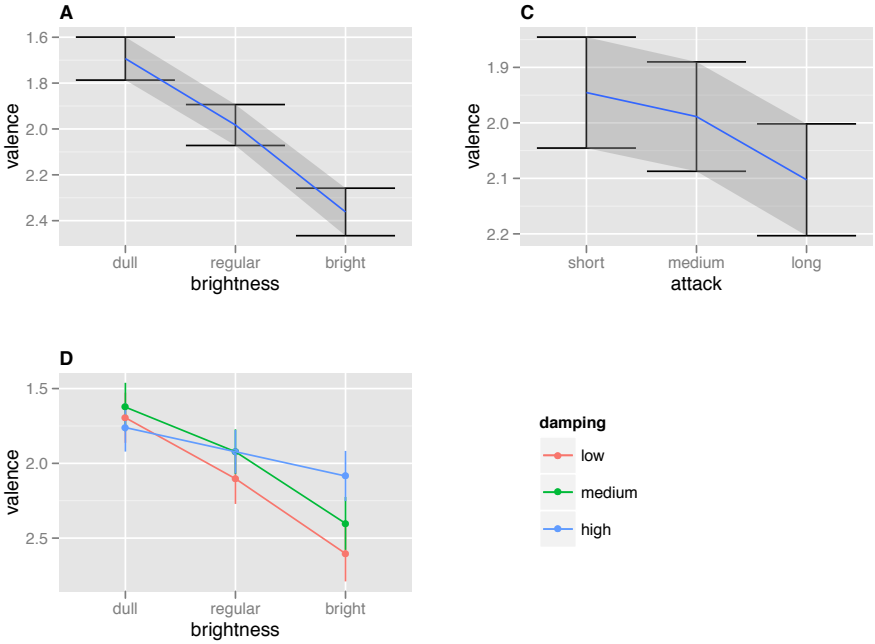


Fig. 11. Effect of timbre parameters (Brightness, Damping and Attack time) on Valence. A) Longer attack time are perceived as more negative **B)** Bright sounds tend to be perceived more negatively than dull sounds **C)** Interaction effects between damping and brightness show that a sound with high damping attenuates the negative valence due to high brightness.

Dominance Follow-up univariate analysis revealed main effects of **Brightness** $F(1.49, 13.45) = 6.55, p < 0.05$ and **Damping** $F(1.05, 10.915) = 4.7, p < 0.05$ on Dominance (Figure 3).

A significant mean difference of -0.743 was found between high and low **brightness**. The brighter the more dominant.

A significant mean difference of 0.33 was found between medium and low **damping** factor. The more damped the less dominant.

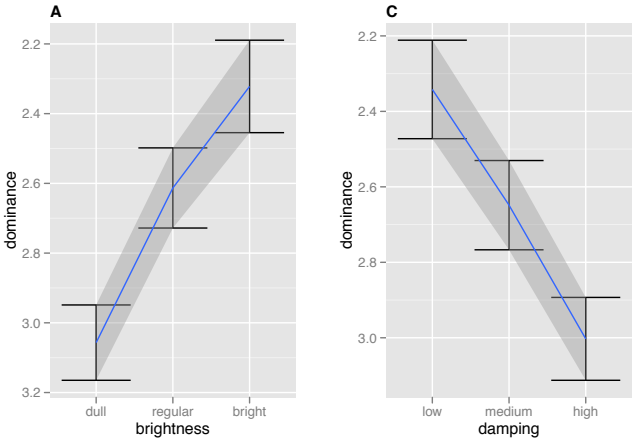


Fig. 12. Effect of timbre parameters (Brightness and Damping) on Dominance. A) Bright sounds are perceived as more dominant than dull sounds **B)** A sound with medium damping is perceived as less dominant than low damping.

4 Conclusions

This study validates the use of the SMuSe as an "affective music engine". The different levels of musical parameters that were experimentally tested evoked significantly different emotional responses. The tendency of minor mode to increase negative valence and of high register to increase arousal (Figure 5) corroborates the results of [12,13], and is complemented by interaction effects (Figure 6). The tendency of short articulation to be more arousing and more negative (Figure 7 and 8) confirms results reported in [14,15,16]. Similarly, higher tempi have a tendency to increase arousal and decrease valence (Figure 7 and 8) are also reported in [14,15,12,13,17,16]. The present study also indicates that higher tempi are perceived as more dominant (Figure 9). Musical sequences that were played louder were found more arousing and more negative (Figure 7 and 8) which is also reported in[14,15,12,13,17,16], but also more dominant (Figure 9). The fact that higher brightness tends to evoke more arousing and negative responses (Figure 10 and 11) has been reported (but in terms of number of harmonics in the spectrum) in [13]. Additionally, brighter sounds are perceived as more dominant (Figure 12). Damped sounds are less arousing and dominant (Figure 10 and 12). Sharp attacks are more arousing and more positive (Figure 10 and 11). Similar results were also reported by [14]. Additionally, this study revealed interesting interaction effects between damping and brightness (Figure 10 and 11).

Most of the studies that investigate the determinants of musical emotion use recordings of musical excerpts as stimuli. In this experiment, we looked at the effect of a well-controlled set of synthetic stimuli (generated by the SMuSe) on the listener's emotional responses. We developed an automated test procedure

that assessed the correlation between a few parameters of musical structure, expressivity and timbre with the self-reported emotional state of the participants. Our results generally corroborated the results of previous meta-analyses [15], which suggests our synthetic system is able to evoke emotional reactions as well as "real" musical recordings. One advantage of such a system for experimental studies though, is that it allows for precise and independent control over the musical parameter space, which can be difficult to obtain, even from professional musicians. Moreover with this synthetic approach, we can precisely quantify the level of the specific musical parameters that led to emotional responses on the scale of arousal, valence and dominance. These results pave the way for an interactive approach to the study of musical emotion, with potential application to interactive sound-based therapies. In the future, a similar synthetic approach could be developed to further investigate the time-varying characteristics of emotional reactions using continuous two-dimensional scales and physiology [18,19].

References

1. L.-O. Lundqvist, F. Carlsson, P. Hilmersson, and P. N. Juslin, "Emotional responses to music: experience, expression, and physiology," *Psychology of Music* **37**(1), pp. 61–90, 2009.
2. S. Le Groux and P. F. M. J. Verschure, *Music Is All Around Us: A Situated Approach to Interactive Music Composition*. Exeter: Imprint Academic, April 2011.
3. S. Le Groux and P. F. M. J. Verschure, "Situated interactive music system: Connecting mind and body through musical interaction," in *Proceedings of the International Computer Music Conference*, Mc Gill University, (Montreal, Canada), August 2009.
4. P. N. Juslin and D. Västfjäll, "Emotional responses to music: the need to consider underlying mechanisms," *Behav Brain Sci* **31**, pp. 559–75; discussion 575–621, Oct 2008.
5. J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology* **39**, pp. 345–356, 1980.
6. P. Lang, "Behavioral treatment and bio-behavioral assessment: computer applications," in *Technology in Mental Health Care Delivery Systems*, J. Sidowski, J. Johnson, and T. Williams, eds., pp. 119–137, 1980.
7. S. Le Groux and P. F. M. J. Verschure, "Emotional responses to the perceptual dimensions of timbre: A pilot study using physically inspired sound synthesis," in *Proceedings of the 7th International Symposium on Computer Music Modeling*, (Malaga, Spain), June 2010.
8. D. Zicarelli, "How I learned to love a program that does nothing," *Computer Music Journal* (26), pp. 44–51, 2002.
9. S. McAdams, S. Winsberg, S. Donnadieu, G. De Soete, and J. Krimphoff, "Perceptual scaling of synthesized musical timbres : Common dimensions, specificities, and latent subject classes," *Psychological Research* **58**, pp. 177–192, 1995.
10. J. Grey, "Multidimensional perceptual scaling of musical timbres," *Journal of the Acoustical Society of America* **61**(5), pp. 1270–1277, 1977.
11. S. Lakatos, "A common perceptual space for harmonic and percussive timbres.," *Perception & Psychophysics* **62**(7), p. 1426, 2000.

12. C. Krumhansl, "An exploratory study of musical emotions and psychophysiology," *Canadian journal of experimental psychology* **51**(4), pp. 336–353, 1997.
13. K. Scherer and J. Oshinsky, "Cue utilization in emotion attribution from auditory stimuli," *Motivation and Emotion* **1**(4), pp. 331–346, 1977.
14. P. Juslin, "Perceived emotional expression in synthesized performances of a short melody: Capturing the listener's judgment policy," *Musicae Scientiae* **1**(2), pp. 225–256, 1997.
15. P. N. Juslin and J. A. Sloboda, eds., *Music and emotion : theory and research*, Oxford University Press, Oxford ; New York, 2001.
16. A. Friberg, R. Bresin, and J. Sundberg, "Overview of the kth rule system for musical performance," *Advances in Cognitive Psychology, Special Issue on Music Performance* **2**(2-3), pp. 145–161, 2006.
17. A. Gabrielsson and E. Lindström, *Music and Emotion - Theory and Research*, ch. The Influence of Musical Structure on Emotional Expression. Series in Affective Science, Oxford University Press, New York, 2001.
18. O. Grewe, F. Nagel, R. Kopiez, and E. Altenm
"uller, "Emotions over time: Synchronicity and development of subjective, physiological, and facial affective reactions to music," *Emotion* **7**(4), pp. 774–788, 2007.
19. E. Schubert, "Modeling perceived emotion with continuous musical features," *Music Perception* **21**(4), pp. 561–585, 2004.

Timing synchronization in string quartet performance: a preliminary study

Marco Marchini[1], Panos Papiotis[1], and Esteban Maestre[2,1] *

[1] Music Technology Group, Universitat Pompeu Fabra, Barcelona
[2] Center for Computer Research in Music and Acoustics, Stanford University, CA
{marco.marchini, panos.papiotis}@upf.edu
esteban@ccrma.stanford.edu

Abstract. This work presents a preliminary study of timing synchronization phenomena in string quartet performance. Accurate timing information extracted from real recordings is used to compare timing deviations in solo and ensemble performance when executing a simple musical passage. Multi-modal data is acquired from real performance and processed towards obtaining note-level segmentation of recorded performances. From such segmentation, a series of timing deviation analyses are carried out at two different temporal levels, focusing on the exploration of significant differences between solo and ensemble performances. This paper briefly introduces, via an initial exploratory study, the experimental framework on which further, more complete analyses are to be carried out with the aim of observing and describing certain synchronization phenomena taking place in ensemble music making.

Keywords: music performance, ensemble performance, synchronization, timing, tempo, string quartet

1 Introduction

Music performance as the act of interpreting, structuring and physically realizing a composition is a highly complex human activity with many facets: physical, acoustic, physiological, psychological, social, artistic, etc. [4]. Trained musicians are able to read and interpret a composition in the form of a music score, which may end up conveying very different emotions depending on how it is performed, i.e., how the content is transformed into musical sound. In fact, it is commonly acknowledged that there is an important part of expression or meaning already borne by the actual piece to be performed, and another part introduced by the performer when freely navigating the space of performance resources (e.g., timing deviations, dynamics modulations, etc.) resulting from a combination of praxis habits and certain constraints imposed by the structure and content of the score [2]. In the search for exploring and understanding the process of music

* The authors would like to thank collaborators from CIRMMT, McGill University for their support in hosting the recordings: Carolina Brum, Erika Donald, Vincent Freour, Marcello Giordano, and Marcelo M. Wanderley.

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

performance as an accessible instance of human cognition, some researchers of a variety of disciplines have tried to approach the challenge by looking at music performance as a goal-directed task, considering such task as driven by the sequence of symbolic events appearing in a music score [11].

Ensemble music performance can be regarded as one of the most closely synchronized activities that human beings engage in (actions coordinated to within small fractions of a second are considered routine even in amateur performance). Unlike speech, musical performance is one of the few expressive activities allowing simultaneous participation. As such, the potential of music as a basis for studying basic principles of non-verbal communication and entrainment of emotion is unparalleled [7]. Studies in computational modeling of music performance have confirmed the widespread consideration of tempo and dynamics as the two most prominent resources available for musicians to convey emotion or expression during performance [14] (e.g., by acting with creative freedom to carve personalized and aesthetically pleasing executions), thus represeting two major dimensions over which to extract relevant information from a performance recording of a certain piece. In agreement with the importance of explicitly considering metric, melodic and harmonic structures of scores when approaching the study of music performance from a computational perspective [1], previous researchers have based their work on pairing musicological charactersitics of musical scores with performance aspects, especially timing and/or dynamics [3, 13, 15, 12]. From these two dimensions, available for ensemble musicians to coordinate and successfully achieve their shared goal, a first clear choice for extracting synchronization-related information from joint performance is to analyze how timing modulations get synchronized in different situations (e.g., solo versus ensemble) and different musical contexts (e.g., by accounting for score structure).

The computational study of timing synchronization among ensemble performers has been approached in the past. A vast literature has been inspired by the concept of "participatory discrepancy" introduced in [6] by Keil. Following the Keil's directions, an objective measure of performer's time discrepancies for several bass players was out carried in [10]. However only a one-way synchronization could be observed since the musicians where recorded playing solo on top of a recorded tape. More interaction paradigms were considered in the work by Goebl and Palmer [5], where the focus was put onto exploring the influence of auditory feedback and musical role (e.g., leadership) on timing (note onsets) and motion (finger and head) synchronization phenomena among duets of pianists. A second relevant example of two-ways auditory/visual feedback is the work by Moore and Chen [9], which pursued micro-timing analyses from arm motion data acquired from two members of a string quartet while performing a relatively difficult, yet thoroughly rehearsed task. Findings of both works showed timing and/or motion synchronization as an essential cue for the exploration of basic social behaviors in coordinated action.

In this paper, we present a preliminary study of timing synchronization phenomena among the four members of string quartet during performance. Timing information is extracted by processing multi-modal data acquired from real

recording (providing a note-level score-performance alignment) of the execution of a simple musical passage that was unknown to the musicians. From annotated note onsets and offsets, a number of timing deviation analyses are carried out at two different temporal levels, focusing on the exploration of significant differences between solo and ensemble performances. Rather than with the aim of presenting a thorough study on the topic, this paper discusses an initial exploration experiment while introducing the framework and methods through which more complete and extended analyses are to be systematically carried out on a large corpus of quartet performance recordings being constructed at the moment.

The rest of the paper is structured as follows. Sec. 2 briefly introduces the general approach we envisioned and employed and explains what type of data we acquire from each experiment and summarizes the techniques used for preprocessing it. We then present some preliminary results on tempo in Sec. 3 and, finally, discuss them in Sec. 4.

2 Experimental framework

As verified by the above literature, the subject of collaborative musical performance is a very complex one. In order to obtain reliable results using computational means, the existence of valid hypotheses is of very high value; for that reason, we are working on an experimental framework which will provide a set of recordings where the studied relationships among the musicians are well defined and unambiguous. The final corpus of music pieces used, which will be detailed next, has been selected and modified using the help of a professional string quartet performer, and will in time be recorded by a number of different quartets.

The corpus is based on an exercise handbook for string quartets[3], intended for improving the "ensemble skills" of the quartet members. The material is divided into six categories, with each category containing a number of short exercises dealing with a different aspect of ensemble performance: *Intonation, Dynamics, Unity of Execution, Rhythm, Phrasing, and Tone production/Timbre.* An exercise consists of a simple, low difficulty score, together with annotations on what is the specific goal that must be achieved by the quartet.

We record the musicians' performance in three experimental conditions; solo (first sight), rehearsal, and ensemble. In the first condition (solo), each musician must perform their part alone without having access to the full ensemble score nor the instructions that accompany the exercise. In this way we wish to eliminate any type of external influence on the performance, be it restrictions imposed by other voices of the ensemble or instructions by the composer that are not in relation to the individual score of the performer. In the second condition (rehearsal) , following the solo recordings of each quartet member, the group of musicians is provided with the full ensemble score plus the composer instructions; they are then left to rehearse the exercise alone until they are able to fulfill

[3] Mogens Heimann - Exercises for the String Quartet.

the requirements of the exercise. In the third condition (ensemble), following the rehearsal, the quartet is finally recorded performing the exercise as a group.

In terms of data acquisition, both audio and motion capture data are recorded for each member of the ensemble; these data streams are synchronized in real time. Audio-wise, the individual signal from each musician is captured through the use of piezoelectric pickups attached to the bridge of the instrument. Instrumental - i.e. sound-producing - gestures such as bow velocity and force are also acquired through the use of a motion capture system, as detailed in [8].

For every recording a semi-automatic, note-level alignment between the performance and the music score is performed using a dynamic programming approach, a variation of the well-known Viterbi algorithm. This approach focuses into three main regions of each note: the note body and two transition segments (onset and offset). Different costs are computed for each segment, using features extracted by the audio (RMS audio energy, Fundamental frequency) as well as the bowing features described above. Finally, the optimal note segmentation is obtained so that a total cost (computed as the sum of the costs corresponding to the complete sequence of note segments) is minimized. This method, which can be seen in more detail in [8], has so far provided robust results that only in few occasions require manual correction. Through this alignment, it is then easy and accurate to extract detailed timing information for each performer. More complicated information such as the dynamics, timbre or articulation of the performance is extracted by combining the audio signal with the instrumental gesture features.

3 Preliminary study and initial results

The objective of the study presented here is to exploit content of some preliminary experiments and formulate new hypothesis to be tested in the next set of experiments. In this article we deal with some results arising from the experiment conducted with the exercise shown in Fig. 1. The exercise consists of an ascending and descending D major scale in thirds. The quartet is divided in two sections (violins in the first, viola and cello in the second) one alternating with the other. Musicians were instructed to play the score as if it was played by one instrument. We did not impose on them further constraints such as to follow a metronome.

For each case we recorded 4 consecutive repetition of the score (Fig. 1). A score alignment has been executed on each of those 8 performances. The analyzed set of data thus consists of 512 aligned onsets (64 per performance). We also derived a *joint-performance alignment* consisting of 128 onsets where for each third chord we compute an onset given by the mean of the individual attacks of the two notes that form that chord. The goal of this exercise is self-evident in the score. The notes within each group have to blend together while allowing the blocks of semiquaver notes formed by each group to slot together in a temporal order. In addition to that, the requirement of achieving a good "unity of execution" means that the parts played by each group have to be connected

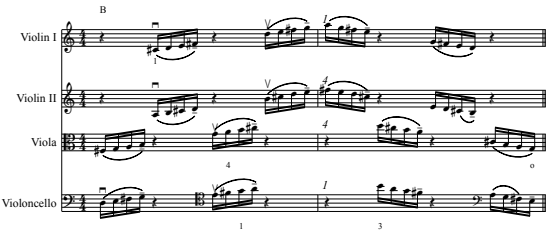


Fig. 1. Score of the exercise employed for the experiment.

to the part of the subsequent section without disruptions in terms of tempo and dynamics. It is also worth to notice that the slurs contained in the scores, by requiring the musicians perform with a certain bow direction might also pose some constraints to the synchronization process.
We divided results of the analysis into *Micro-* and *Macro-* tempo results. We consider *Macro-tempo* as the tempo experienced by a listener in a relatively long region of time, it can also change in time but rather slowly. *Micro-Tempo* comprises of slight anticipations of note events followed by a deferral of the subsequent events in a way that the result does not contribute to macro-tempo changes.

3.1 Macro-tempo

By assuming, for a moment, the tempo to be constant when no onset occurs we derive a bpm step function defined to be the corresponding beat per minute value of each duration. This curve is noisy due to the differences in duration of the notes. In order to remove high frequency content and derive an overall tempo behavior we convolve it with a gaussian curve of variance $\sigma > 0$. From

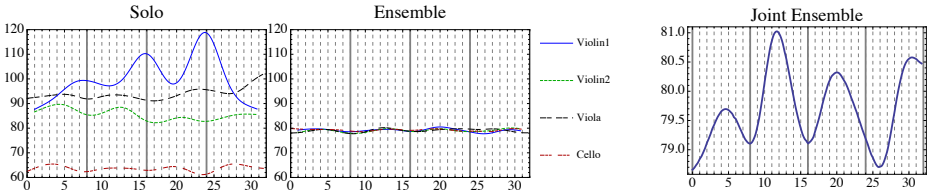
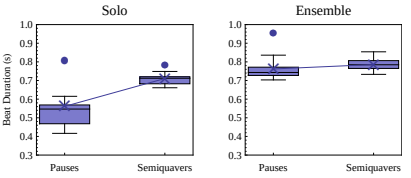


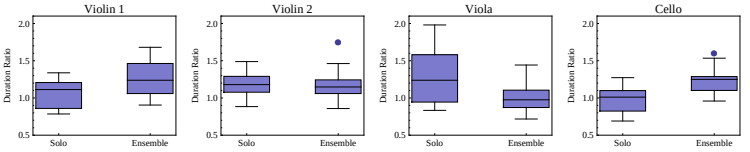
Fig. 2. Individual tempo curve of the four instruments for the solo (left) and the ensemble case (center). Vertical grid lines mark the boundaries of the repetitions (full line) and the beat start time (dashed line).The tempo curve of the joint ensemble performance is shown on the last plot (right).

the tempo curves derived for the solo/ensemble case we find that not only the mean tempo of the four musicians becomes the same, but also the variance of

the tempo curve gets significantly smaller in the ensemble case. We can interpret this result both as an indicator that the freedom of the musicians gets restricted and as a result of the collaborative way in which the tempo is jointly shaped. Fig. 2 shows tempo derived with $\sigma = 1.67$ for the solo/ensemble case. As it is clear from the plots, the individual tempo curves contract to the same tempo when the musicians play together.



(a) Beat duration for the first violinist solo vs ensemble.



(b) SQDR of solo vs ensemble

Fig. 3. Box-and-whisker diagrams showing some results in micro-tempo.

In the solo excerpts we found a relationship between the alternation of pauses and semiquavers of each single voice and the corresponding duration of the beat. The tempo was kept differently by the musicians in the case of a pause then in the case of semiquavers. Also in this case we found differences from the ensemble recordings case where the discrepancy between pause and semiquaver duration gets smaller because of the interdependence among musicians. The results of this analysis are shown in Fig. 3(a) by box-and-whisker diagrams. A t-test shows a significant difference between pauses and semiquavers duration in the cases of solo violin 1 and solo viola. In the ensemble case only the viola is found to play pauses consistently shorter than semiquavers although the difference was small[4]. Regarding the variances, a χ^2-test at a significance level of 5% could find disjoint confidence intervals only for the first violin in the solo and the cello in the ensemble. In the remaining cases the amount of variation in duration across the pauses did not differ from the one of the notes.

Fig. 3.1 shows the same tempo analysis for the joint ensemble performance. The most evident feature of this tempo curve is its relationship with the repeti-

[4] The difference between the mean duration of the four notes groups and that of the pauses was just 47 ms. This does not necessarily mean that the viola was not synchronized with the rest but it might mean that he was slightly anticipating the the others' first note onset and/or deferring the last note offset in the group of notes.

tion structure of the exercise. In fact, while remaining relatively constant (just slightly increasing through the performance) the performance tempo was indeed oscillating by speeding up in the center of the repetition and slowing down towards its boundaries. In table 3.1 the correlation with the pitch curve[5] is shown for all the cases. In the joint ensemble performance the correlation is 0.56. This confirms the overall tendency of the performance to speed up at higher pitches[6]. Despite the fact that we only have recoded few repetitions, this value of correlation is highly improbable to arise by chance. To quantify the significance we have used an empirical (Monte-Carlo) method. We generated sample random performances by perturbing in different ways the score with a gaussian noise to each note onset time. For each random performance we have performed the same macro-tempo analysis as the one performed on the real performance. Five groups of 2000 random performances were generated of respectively standard deviations 0.1, 2.5, 5, 10 and 25 ms. We then yield a value of pitch correlation for the tempo curve produced by each perturbed score and estimate variance and mean of the correlation. Assuming it to be a gaussian distribution we obtain an empirical p-value[7] for each noise amplitude. The resulting p-values are shown in Fig. 5 and, as you can see, the p-value is bounded by 2.5%. Remarkably, an increase of error variance σ^2 yields a decrease of p-values and not the other way around. Thus, it is even less likely to get a big correlation by adding a bigger noise than by adding a small one. In conclusion a confidence level of 98% can be considered to hold in all cases.

We have to notice that the excursion of the tempo curve is smaller than the just-noticeable-difference (JND). This means that the musicians are not aware of this fluctuations of tempo. Moreover, we can not still distinguish if this mechanism is directly related to repetition structure, pitch or to some more complex unconscious mechanism governing the performance.

	Solo		Ensemble		Joint Ensemble	
	Corr	Cov	Corr	Cov	Corr	Cov
Violin 1	-0.54	-33.65	0.59	2.87	0.56	2.78
Violin 2	0.5	8.5	0.75	3.34		
Viola	0.05	0.92	0.59	3.01		
Cello	0.72	6.1	0.35	1.01		

Fig. 4. Correlation of pitch with joint ensemble tempo curve.

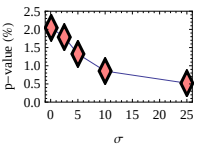


Fig. 5. p-value for different variances of gaussian noise for the empirical significance test.

[5] The pitch curve has values in number of semitones and has been constructed by taking the higher pitched note of each chord

[6] This is predicted by the well-known phrase-arch rule of Friberg et al. It is thus probably unrelated to pitch, and occurs only because the high pitches are in the middle of the phrase. However the performance we are analyzing here, far from being expressive, is just an exercise scale.

[7] The empirical probability of having a correlation as high as the one measured for the real performance (0.56).

3.2 Micro-tempo

Whereas macro-tempo can be related to global properties of the performance such as phrases or repetition patterns, micro-tempo is usually related to incidental local characteristics of the score.

At a shorter micro tempo scale, we found a consistent relation between the duration of each semiquaver and the position it occupies within groups of 4 semiquavers. An ANOVA test could confirm at a significance level lower than 1% the effect of metrical position on the joint-performance.

Differences have been found also when comparing the solo performance with the ensemble performance. Since the general tendency is to play the first note of the group longer than the second we have focused, for the sake of simplicity, on the ratio between the duration of the first semiquaver duration and the second (SQDR). This simplification also enables us to compare the solo case with the ensemble case since the ratio is not directly dependent on tempo. Remarkably, we could prove at a significance level lower than 2% the effect of the two scenarios to the SQDR for first Violin, Viola and Cello. We can thus report an overall tendency to exaggerate the agogic accent of consecutive strong-weak semiquaver couples in the ensemble case respect to the solo. Whereas the second violin keeps maintaining a positive SQDR of 1.19 in both the cases, the first violinist and the cello increase theirs from 1.07 to 1.27 and from 1.0 to 1.24 respectively. Despite this general tendency, a different behavior was measured for the viola which was decreasing its SQDR from 1.29 to 1.02.

A further analysis of the precedence of the onset times seems to explain the different micro-tempo results of the musicians in the ensemble case. Analyzing the attack time of the musicians having synchronized notes we found out that the attacks of the cello were preceding the ones of the viola by a mean of 8 ms, and the first violin was preceding the second by 13 ms. Musicians employing an higher SQDR are thus also anticipating their partner on the average. This suggest that the use of contrast in successive notes could be used as a mean of communication between the musicians to better control the synchronization.

4 Discussion and future work

We have presented an experimental framework through which we assign the musicians of a string quartet the task of playing specifically chosen exercises after a brief rehearsal period. In this context we have shown a set of preliminary results on timing synchronization phenomena observing the differences between musicians playing alone or in ensemble.

In the macro-tempo and at the beat level we have observed broad reduction of the mean bpm and its total variation in each single instrument. This confirms the hypothesis that the constraints that musicians are required to follow end in favoring a more controlled execution. In the joint ensemble performance we have then detected a consistent correlation of the bpm with the phrase structure of the repetition. Despite the fact that the excursion was here within the JND for

tempo changes we have shown that this behavior is unlikely to happen by chance. However, more experiments should be carried out to check if this behavior arises because of the repetition structure, because of the pitch contour or for more complex reasons.

The analysis micro-tempo, on the other hand, was pointing out generally a bigger variance between short contiguous notes in the ensemble than in the solo. By also looking at the precedence of onset attack time between musicians we have formulated the hypothesis that a bigger contrast between contiguous short note duration might be used by leaders to maximize the communication with the other musicians or improve the synchronization. This hypothesis should be taken into account systematically to design further experiments.

References

1. J. Beran and G. Mazzola. Analyzing musical structure and performance-a statistical approach. *Statistical Science*, 14(1):pp. 47–79, 1999.
2. D. Deutsch, editor. *The psychology of music.* Academic Press, 2nd edition, 1998.
3. A. Friberg. Generative rules for music performance: A formal description of a rule system. *Computer Music Journal*, 15(2):56–71, 1991.
4. A. Gabrielsson. *The performance of Music.* Academic Press., 1999.
5. W. Goebl and C. Palmer. Synchronization of timing and motion among performing musicians. *Music Perception*, 26(5):427–438, 2009.
6. C. Keil. Participatory discrepancies and the power of music. *Cultural Anthropology*, 2(3):275–283, 1987.
7. P. Keller. Joint action in music performance. *Emerging Communication*, 10:205, 2008.
8. E. Maestre. *Modeling instrumental gestures: an analysis/synthesis framework for violin bowing.* PhD thesis, Universitat Pompeu Fabra, Barcelona, Spain, Novembre 2009.
9. G. P. Moore and J. Chen. Timings and interactions of skilled musicians. *Biol. Cybern.*, 103:401–414, November 2010.
10. J. A. Prögler. Searching for swing: Participatory discrepancies in the jazz rhythm section. *Ethnomusicology*, 39(1):pp. 21–54, 1995.
11. H. Purwins and D. R. Hardoon. Trends and perspectives in music cognition research and technology. *Connection Science*, 21:85–88, June 2009.
12. R. Ramirez, A. Hazan, E. Maestre, and X. Serra. A data mining approach to expressive music performance modeling. *Multimedia Data Mining and Knowledge Discovery*, pages 362—380, 2007.
13. N. Todd. The dynamics of dynamics: A model of musical expression. *The Journal of the Acoustical Society of America*, 91(6):3540–3550, 1992.
14. G. Widmer and W. Goebl. Computational models of expressive music performance: The state of the art. *Journal of New Music Research*, 33(3):203–216, 2004.
15. G. Widmer and A. Tobudic. Playing mozart by analogy: Learning multi-level timing and dynamics strategies. *Journal of New Music Research*, 32(3):259–268, 2003.

Predicting Time-Varying Musical Emotion Distributions from Multi-Track Audio

Jeffrey Scott, Erik M. Schmidt, Matthew Prockup, Brandon Morton, and Youngmoo E. Kim

Music and Entertainment Technology Laboratory (MET-lab)
Electrical and Computer Engineering, Drexel University
{jjscott, eschmidt, mprockup, bmorton, ykim}@drexel.edu

Abstract. Music exists primarily as a medium for the expression of emotions, but quantifying such emotional content empirically proves a very difficult task. Myriad features comprise emotion, and as such music theory provides no rigorous foundation for analysis (e.g. key, mode, tempo, harmony, timbre, and loudness all play some roll), and the weight of individual musical features may vary due to the expressiveness of different performers. In previous work, we have shown that the ambiguities of emotions make the determination of a single, unequivocal response label for the mood of a piece of music unrealistic, and we have instead chosen to model human response labels to music in the arousal-valence (A-V) representation of affect as a *stochastic distribution*. Using multi-track sources, we seek to better understand these distributions by analyzing our content at the performer level for different instruments, thus allowing the use of instrument-level features and the ability to isolate affect as a result of different performers. Following from the time-varying nature of music, we analyze 30-second clips on one-second intervals, investigating several regression techniques for the automatic parameterization of emotion-space distributions from acoustic data. We compare the results of the individual instruments to the predictions from the entire instrument mixture as well as ensemble methods used to combine the individual regressors from the separate instruments.

Keywords: emotion, mood, machine learning, regression, music, multi-track

1 Introduction

There has been a growing interest in the music information retrieval (Music-IR) research community gravitating towards methods to model and predict musical emotion using both content based and semantic methods [1]. It is natural for humans to organize music in terms of emotional associations, and the recent explosion of vast and easily accessible music libraries has created high demand for automated tools for cataloging, classifying and exploring large volumes of music content. Crowdsourcing methods provide very promising results, but do not perform well outside of music that is highly-popular, and therefore leave much

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

to be desired given the long-tailed distribution of music popularity. The recent surge of investigations applying content-based methods to model and predict emotional affect have generally focused on combining several feature domains (e.g. loudness, timbre, harmony, rhythm), in some cases as many as possible, and performing dimensionality reduction techniques such as principal component analysis (PCA). While using these methods may in many cases provide enhanced classification performance, they provide little help in understanding the contribution of these features to musical emotion.

In this paper, we employ multi-track sources for music emotion recognition, allowing us to extract instrument-level acoustic features while avoiding corruption that would usually occur as a result of noise induced by the other instruments. The perceptual nature of musical emotion necessarily requires supervised machine learning, and we therefore collect time-varying ground truth data for all of our multi-track files. As in previous work, we collect data via a Mechanical Turk human intelligence task (HIT) where participants are paid to provide time-varying annotations in arousal-valence (A-V) model of affect, where valence indicates positive vs negative emotion, and arousal indicates emotional intensity [2]. In this initial investigation we obtain these annotations on our full multi-track audio files, thus framing the task as predicting the mixed emotion from the individual instrument sources. Furthermore, we model our collected A-V data for each moment in a song as a *stochastic distribution*, and find that the labels can be well represented as a two-dimensional A-V Gaussian distribution.

In isolating specific instruments we gain the ability to extract specific acoustic features targeted at each instrument, allowing us to find the most informative domain for each. In addition, we also isolate specific performers, potentially allowing us to take into account performer-level affect as a result of musical expression. We build upon our previous work modeling time-varying emotion-space distributions, and seek to develop new models to best combine this multi-track data [3–5]. We investigate multiple methods for automatically parameterizing an A-V Gaussian distribution, effectively creating functional mappings from acoustic features directly to emotion space distribution parameters.

2 Background

Prior work in modeling musical emotion has explored content based and semantic methods as well as combinations of both models [1]. Much of the work in content based methods focuses on training supervised machine learning models to predict classes of emotion, such as happy, joyful, sad or depressed. Several works also attempt to classify songs into discretized regions of the arousal-valence mood space [6–8].

In addition to classification, several authors have successfully applied regression methods to project from high dimensional acoustic feature vectors directly into the two dimensional A-V space [9, 8]. To our knowledge, no one has attempted to leverage the separate audio streams available in multi-track recordings to enhance emotion prediction using content based methods.

3 Dataset

We selected 50 songs spanning 50 unique artists from the RockBand® game and created five monaural stem files for each song. This is the same dataset (plus 2 additional songs) that we used in a previous paper for performing analyses on multi-track data[10, 11]. A stem may contain one or more instruments from a single instrument class. For example, the vocal track may have one lead voice or a lead and harmony or even several harmonies as well as doubles of those harmonies. Each stem only contains one instrument class (i.e. bass, drums, vocals) excepting the backup track which can contain audio from more than one instrument class. For each song there are a total of six audio files - backup, bass, drums, guitar, vocals and the full mix, which is a linear combination of the individual instruments.

To label the data, we employed an annotation process based on the MoodSwings game outlined in [2]. We used Amazon's Mechanical Turk and rejected the data of users who did not pass the verification criteria of consistent labeling on the same song and similarity to expert annotations. For the 50 songs in our corpus there is an average of 18.48 ± 3.05 labels for each second with a maximum of 25 and a minimum of 12. A 40 second clip was selected for each song and the data of the first 10 seconds was discarded due to the time it takes a user to decide on the emotional content of the song [12]. As a result, we are using 30 second clips for our time varying prediction of musical emotion distributions.

4 Experiments

The experiments we perform are similar in scope to those presented in a previous paper which utilized a different dataset [4]. This allows us to verify that we attain comparable results using instrument mixtures and provides a baseline to compare the results from the audio content of individual instruments.

4.1 Overview

Acoustic features are extracted from each of the five individual instrument files as well as the final mix and are described in more detail in Section 4.2. We use linear regression to calculate the projection from the feature domain of each track to the parameters of the Gaussian distribution that models the labels at a given time.

$$[f_1^{(t)} \cdots f_m^{(t)}]\, \mathbf{W}_t = [\mu_v^{(t)}\ \mu_a^{(t)}\ \mathbf{\Sigma}_{11}^{(t)}\ \mathbf{\Sigma}_{12}^{(t)}\ \mathbf{\Sigma}_{22}^{(t)}] \tag{1}$$

Here $[f_1^{(t)} \cdots f_t^{(t)}]$ are the acoustic features, $\mathbf{W}_t$ is the projection matrix, μ_a and μ_v are the means of the arousal and valence dimensions, respectively, and $\mathbf{\Sigma}$ is the 2×2 covariance matrix. For an unknown song, $\mathbf{W}_t$ is used to predict the distribution parameters in the A-V space from the features for track t. The regressor for each track can be used on its own to predict A-V means and covariances.

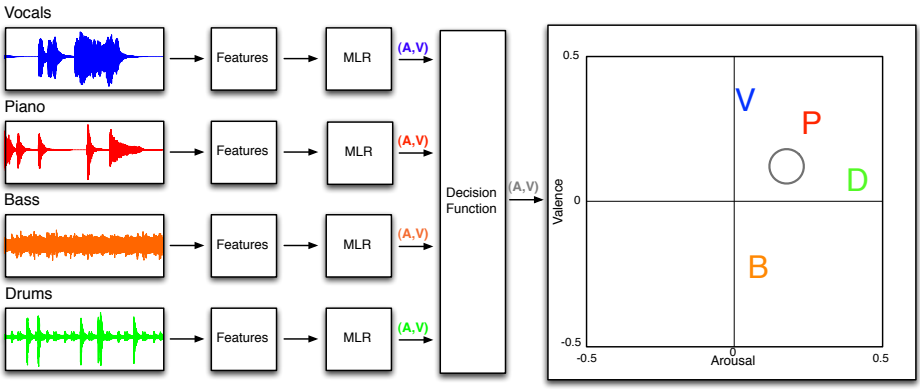


Fig. 1: Acoustic features are computed on each individual instrument file and a regression matrix is computed to project from features to a distribution in the A-V space. A different distribution is computed for each instrument (B/D/P/V) and the mean of the distribution parameters (gray circle) is used as the final A-V distribution.

We also investigate combinations of the individual regressors to reduce the error produced by a single instrument model. In these cases, the final prediction is a weighted combination of the predictions from each individual regressor

$$\boldsymbol{\theta} = \sum_{k=1}^{K} \pi_K \theta_K \tag{2}$$

where $\theta = [\mu_v \ \mu_a \ \boldsymbol{\Sigma}_{11} \ \boldsymbol{\Sigma}_{12} \ \boldsymbol{\Sigma}_{22}]$ and π_k is the mixture coefficient for each regressor. In this paper, we try the simplest case which averages the predicted distribution parameters to produce the final distribution parameter vector. Figure 1 depicts the test process for an unknown song.

Having a small dataset of only 50 songs, we perform leave-one-out cross validation (LOOCV), training on 49 songs and testing on the remaining song. This process is repeated until every song has been used as a test song.

4.2 Acoustic Features

We investigate the performance of a variety of acoustic features that are typically used throughout the music information retrieval (Music-IR) community including MFCCs, chroma, spectrum statistics and spectral contrast features. The audio files are down-sampled to 22050 Hz and the features are aggregated over one second windows to align with the second by second labels attained from the annotation task. Table 1 lists the features used in our experiments [13–16].

Feature	Description
MFCC	Mel-Frequency Cepstral Coefficients (20 dimensions)
Chroma Autocorrelation	The autocorrelation of the 12 dimensional chroma vector
Spectral Contrast	Energy in spectral peaks and valleys
Statistical Spectrum Descriptors	Statistics of the spectrum (spectral shape)

Table 1: Acoustic features used in the experiments.

5 Results

We perform experiments using the audio of individual instruments, the full instrument mixture and combinations of the individual instruments. We also compare the results of using different features for each track.

Table 2 shows the results for the regressors trained on individual instruments. The mean average error is the average euclidean distance of the predicted mean of the distribution from the true mean of the distribution across all cross validation folds. Since we are modeling distributions and not just singular A-V coordinates, we also compute the one-way Kullback-Liebler (KL) Divergence from the projected distribution to the true distribution of the collected A-V labels. The table shows the average KL divergence for each regressor averaged across all cross validation folds. We observe that the best regressor for bass, drums and vocals is attained using spectral contrast features and the best regressor for the backup and drum tracks is computed using spectral shape features. It is notable that chroma features perform particularly poor in terms of KL divergence but are only slightly worse than the other features at predicting the means of the distribution.

We also consider combinations of regressors which are detailed in Table 3. The 'Best Single' row shows the best performing single regressor in terms of A-V mean prediction using each feature. The second row in the table includes the results of averaging the predicted distribution parameters for all five individual instrument models for the given feature. Lastly, 'Final Mix' lists the average distance between the predicted and true A-V mean when projecting from features computed on the final mixed track. We note that averaging the models improves performance for all of the best single models excepting the spectral contrast feature. Comparing the averaged models to the prediction from the final mix, the averaged single instrument regressors perform better for MFCCs and spectral shape features but do not perform as well as the final mixes when using chroma or spectral contrast features.

In Figure 2 we see examples of both the predicted and actual distributions for a 30 second clip from the song *Hysteria* by Muse. Both the true and estimated distributions get darker over time as do the data points of the individual users. The predictions for the individual instruments (a-e) are shown along with the average of the predictions for all the instruments (f).

Feature	Instrument	Average Mean Distance	Average KL Divergence
MFCC	Backup	0.152 ± 0.083	1.89 ± 2.34
	Bass	0.141 ± 0.070	1.26 ± 1.29
	Drums	0.140 ± 0.075	1.17 ± 1.52
	Guitar	0.133 ± 0.066	1.22 ± 1.40
	Vocals	0.134 ± 0.071	1.41 ± 1.81
Spectral Contrast	Backup	0.145 ± 0.125	1.86 ± 5.93
	Bass	$\mathbf{0.140 \pm 0.076}$	$\mathbf{1.21 \pm 1.38}$
	Drums	0.139 ± 0.071	1.20 ± 1.88
	Guitar	$\mathbf{0.125 \pm 0.063}$	$\mathbf{1.06 \pm 1.42}$
	Vocals	$\mathbf{0.129 \pm 0.065}$	$\mathbf{1.00 \pm 1.32}$
Spectral Shape	Backup	$\mathbf{0.132 \pm 0.068}$	$\mathbf{1.25 \pm 1.91}$
	Bass	0.142 ± 0.071	1.31 ± 1.63
	Drums	$\mathbf{0.131 \pm 0.072}$	$\mathbf{1.03 \pm 1.38}$
	Guitar	0.134 ± 0.063	1.12 ± 1.42
	Vocals	0.133 ± 0.067	1.12 ± 1.47
Chroma	Backup	0.153 ± 0.084	10.85 ± 15.6
	Bass	0.159 ± 0.084	5.35 ± 6.13
	Drums	0.162 ± 0.089	2.87 ± 3.01
	Guitar	0.147 ± 0.074	2.66 ± 4.33
	Vocals	0.154 ± 0.078	5.99 ± 10.4

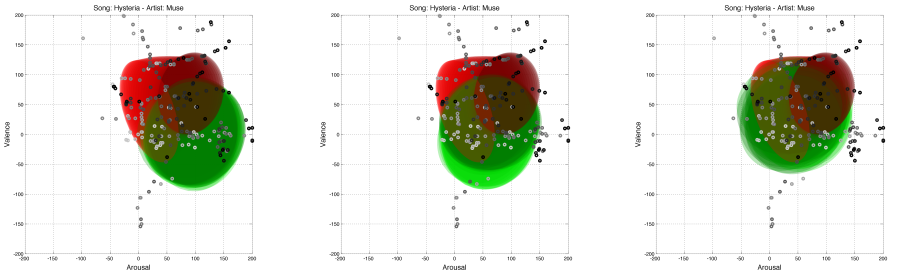
Table 2: Mean average error between actual and predicted means in the A-V coordinate space as well as Kullback-Leibler (KL) divergence between actual and predicted distributions. The value of the best performing feature for each instrument is in bold.

6 Discussion

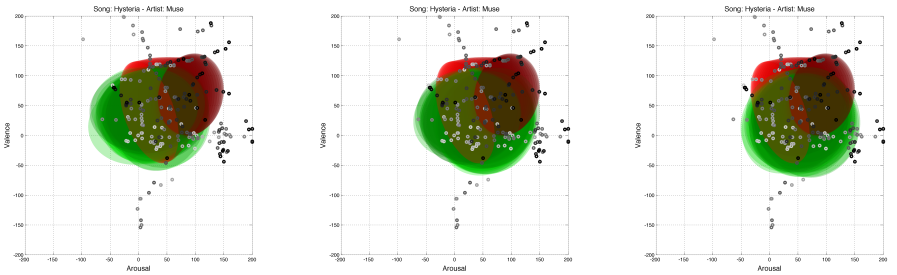
In this initial work we demonstrate the potential of utilizing multi-track representations of songs for modeling and predicting time varying musical emotion distributions. We achieved performance on par with what we have shown previously with a different corpus using similar techniques and a simple averaging of a set of regressors trained on individual instruments. Using more advanced techniques to determine the optimal combinations and weights of instruments and features could provide significant performance gains compared to averaging the output of all the models. There are a variety of ensemble methods for regres-

	Features			
	Chroma	Contrast	MFCC	Shape
Best Single	0.147 ± 0.074	0.125 ± 0.063	0.133 ± 0.066	0.131 ± 0.072
Avg Models	0.142 ± 0.075	0.126 ± 0.066	$\mathbf{0.124 \pm 0.061}$	$\mathbf{0.129 \pm 0.064}$
Final Mix	$\mathbf{0.141 \pm 0.073}$	$\mathbf{0.124 \pm 0.066}$	0.129 ± 0.069	0.132 ± 0.066

Table 3: Results from different combinations of single instrument regressors



(a) Bass (b) Backup (c) Guitar



(d) Drums (e) Vocal (f) Averaged Prediction

Fig. 2: Actual (red) and predicted (green) distributions for *Hysteria* by Muse. The color of the distribution gets darker over time as does the color of the individual data points.

sion that would be applicable to learning better feature and model combinations for regression in the A-V space. We hope to infer, from the results of such experiments, whether certain instruments contribute more to invoking emotional responses from humans.

The results shown in these experiments are encouraging, especially in the performance gains in the case of the MFCC features. An interesting result is that each individual instrument spectral contrast prediction performs better than that of MFCCs, but the MFCC multi-track combination is the top performer equal with spectral contrast on the full mix. This result highlights that the highest performing feature on a single track might not be the same one that offers the most new information to the aggregate track prediction. As a result, in future work we plan to investigate feature selection for this application, performing a number of experiments with different acoustic feature combinations to determine the best acoustic feature for each instrument in the multi-track prediction system.

References

1. Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull, "Music emotion recognition: A state of the art review," in *ISMIR*, Utrecht, Netherlands, 2010.
2. J. Speck, E. Schmidt, and B. Morton, "A comparitive study of collaborative vs. traditional musical mood annotation," in *ISMIR*, Miami, FL, 2011.
3. E. M. Schmidt, D. Turnbull, and Y. E. Kim, "Feature selection for content-based, time-varying musical emotion regression," in *ACM MIR*, Philadelphia, PA, 2010.
4. E. M. Schmidt and Y. E. Kim, "Prediction of time-varying musical mood distributions from audio," in *ISMIR*, Utrecht, Netherlands, 2010.
5. ——, "Prediction of time-varying musical mood distributions using Kalman filtering," in *IEEE ICMLA*, Washinton, D.C., 2010.
6. L. Lu, D. Liu, and H. J. Zhang, "Automatic mood detection and tracking of music audio signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 5–18, 2006.
7. K. Bischoff, C. S. Firan, R. Paiu, W. Nejdl, C. Laurier, and M. Sordo, "Music mood and theme classification-a hybrid approach," in *Proceedings of the 10th International Society for Music Information Conference*, Kobe, Japan, 2009.
8. B. Han, S. Rho, R. B. Dannenberg, and E. Hwang, "Smers: Music emotion recognition using support vector regression," in *ISMIR*, Kobe, Japan, 2009.
9. H. Chen and Y. Yang, "Prediction of the distribution of perceived music emotions using discrete samples," *IEEE TASLP*, no. 99, 2011.
10. J. Scott, M. Prockup, E. M. Schmidt, and Y. E. Kim, "Automatic multi-track mixing using linear dynamical systems," in *SMPC*, Padova, Italy, 2011.
11. J. Scott and Y. E. Kim, "Analysis of acoustice features for automated multi-track mixing," in *ISMIR*, Miami, Florida, 2011.
12. B. G. Morton, J. A. Speck, E. M. Schmidt, and Y. E. Kim, "Improving music emotion labeling using human computation," in *HCOMP '10: Proc. of the ACM SIGKDD Workshop on Human Computation*, Washinton, D.C., 2010.
13. D. Jiang, L. Lu, H. Zhang, J. Tao, and L. Cai, "Music type classification by spectral contrast feature," in *Proc. Intl. Conf. on Multimedia and Expo*, vol. 1, 2002, pp. 113–116.
14. G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *Speech and Audio Processing, IEEE Transactions on*, vol. 10, no. 5, pp. 293–302, 2002.
15. S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE TASSP*, vol. 28, no. 4, 1980.
16. T. Fujishima, "Realtime chord recognition of musical sound: a system using common lisp music." in *Proc. of the Intl. Computer Music Conf.*, 1999.

Codebook Design Using Simulated Annealing Algorithm for Vector Quantization of Line Spectrum Pairs

Fatiha Merazka

Telecommunications Department
University of science & technology Houari Boumediene
P.O.Box 32 El Alia 16111 Bab Ezzouar, Algiers
Algeria

fmerazka@usthb.dz

Abstract. An important issue in vector quantization (VQ) is the design of the codebook. The standard method for codebook design has been the generalized Lloyd algorithm (GLA) and Lind, Buzo and Gray (LBG) algorithm. Theses algorithms can get stuck in suboptimal codebooks due to the presence of several locally minimum distortion values. Simulated annealing (SA) is an optimization procedure that uses randomness to escape local minima in its search for a globally minimum state. In this paper, we propose a method of applying simulated annealing to VQ codebook design problem. The results presented for speech samples represented by line spectrum Pairs (LSP) indicate that the resulting design with simulated annealing are better compared to GLA and LBG algorithms.

Keywords: LSP, Simulated Annealing, GLA, LBG, MSE, SNR

1 Introduction

VQ has become a powerful tool and its application has been frequently reported in the speech and image coding literature [1-3]. The basic definition of a vector quantizer Q of dimension n and size K is a mapping of a vector from n-dimensional Euclidean space, R^n to a finite set, C, containing K reproduction code-vectors [1]:

$$Q: R^n \to C, \tag{1}$$

where $C = \{y_i : i \in I\}$ and $y_i \in R^n$ [1]. Associated with each reproduction code-vector is a partition of R^n, called a region or cell, $S = \{S_i : i \in I\}$ [4]. The most popular form of VQ is the nearest neighbor VQ, where for each input source vector, x, a

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

search is done through the entire codebook to find the nearest code-vector, y_i, which has the minimum distance [5]:

$$y_i = Q(x) \tag{2}$$

$$\text{if } d(x, y_i) < d(x, y_j) \quad for \quad i \neq j \tag{3}$$

where $d(x, y)$ is a distance measure between the vectors, x and y. The mean squared error (*mse*) is used as the distance measure. Depending on the coding application, other more meaningful distance measures may be used such as the the Mahalanobis distance [6], Itakura–Saito distance [7], or other perceptually-weighted distance measures [8].

If the dimension of the vectors is n and a codebook of K code-vectors is used, each vector will be represented as a binary code of length $\lceil \log_2 K \rceil$ bits. Hence the bitrate of the vector quantizer is given by $\frac{1}{n}\lceil \log_2 K \rceil$ bits/sample [5].

Codebook design is the key problem of VQ and the generated codebook has more effect on the compression performance. The most widely used technique to create codebooks is a generalized Lloyd algorithm (GLA)[9], which is an iterative descent technique where an initial codebook is continually refined so that each iteration reduces the distortion involved in coding a given training set. The GLA algorithm provides no guarantee of optimality; a locally optimal solution may be obtained. The Linde–Buzo–Gray (LBG) algorithm [10] is an extension of the iterative Lloyd method [9], for use in VQ design. Because the LBG algorithm is not a variational technique, it can be used for cases where: the probability distribution of the data is not known a priori; we are only given a large set of training vectors; and the source is assumed to be ergodic [10]. The LBG algorithm involves refining an initial set of reproduction code-vectors using the Lloyd conditions [11], based on the given training vectors. The iterative procedure is stopped after the change in distortion becomes negligible.

Research efforts in codebook design have been concentrated in two directions: to generate a better codebook that approaches the global optimal solution, and to reduce the computational complexity.

All of the above algorithms have a local minimum problem. That is, the codebook guarantees local minimum distortion, but not global minimum distortion. To solve this problem, simulated annealing algorithms applied to image coding [12-14] have been proposed. Also, the method of using different initial points to find different codebooks, and then selecting the least distortion codebook as the final codebook, has been investigated. These last two methods can improve the codebook, but they increase the complexity significantly, and they cannot guarantee global optimality.

Competitive learning has also been applied to codebook design [15-20]. Codebook design algorithms based on evolutionary computation are new methods. In the design of the VQ codebook, the genetic algorithm (GA) is a random optimization algorithm based on the process of biological evolution by natural selection and genetic variation. GA has strong global search ability, but a weak local optimum capacity and slow convergence rate. It has advantages of easy use, universality and wide range of application [21-24].

Research on codebook design for VQ using simulated annealing has spanned over twenty years. Most work was focused on codebooks design for image coding. Less attention was paid to codebook design for speech signals.

This paper explores the application of SA algorithm to design codebooks for split VQ (SVQ) of line spectrum pairs (LSP) parameters of speech signals and compare them with GLA and LBG, since these two algorithms remain used most often for developing codebooks [25,26].

Our motivation for the use of LSP coefficient, to represent speech, is due to the fact that in many speech coders, the parameters of the all-zero predictor filter or the corresponding all-pole synthesis filter are coded and sent as part of the information stream [27]-[30].. Recently, there has been a growing interest in the use of (LSP's) to code the filter parameters for linear predictive coding (LPC) of speech

LSP's are an alternative to the direct form predictor coefficients or the lattice form reflection coefficients for representing the filter response. The direct form coefficient representation of the LPC filters is not conducive to efficient quantization. Instead, nonlinear functions of the reflection coefficients (e.g., log-area ratio or inverse sine of the reflection coefficient) are often used as transmission parameters [31]. These parameters are preferable because they have a relatively low spectral sensitivity.

LSP's are an alternate parameterization of the filter with a one-to-one correspondence with the direct form predictor coefficients. The concept of an LSP was introduced by Itakura [32]. LSP's encode speech spectral information more efficiently than other transmission parameters [28,33].

We have opted for the quantization of the LSPs by SVQ. Our choice is justify by the fact that VQ provides greater quantization efficiency than the scalar quantization due to the high correlation between neighboring spectral lines and the intuitive spectral interpolation [1]. Moreover, and in order to make VQ practical for large dimension and high bitrates, a structure can be imposed on the codebook to decrease the search complexity and/or storage requirements. One way of achieving this is to use decompose the codebook into a Cartesian product of smaller codebooks [1,34].

We will apply a 3-3-4 SVQ at 24 bits/frame to test our codebooks design. SVQ was first introduced by Paliwal and Atal [28,35] for quantization of line spectrum frequencies (LSF) in narrowband CELP speech coders and is used in the adaptive multirate narrowband (AMR-NB) codec [36]. SVQ is also used for quantizing Mel frequency-warped cepstral coefficients (MFCCs) in the ETSI distributed speech recognition (DSR) standard [37].

In all cases, the codebook is designed to minimize the mean squared error *mse*, defined by:

$$mse = \frac{1}{k}\sum_{i=0}^{k-1}\frac{1}{n}\left\|(x_i - y_i)\right\|^2 \tag{4}$$

where x_i is the *i*th input sample and y_i is the *i*th output codeword, n is the dimension of the vectors and k is the size of training sequence.

The signal to noise (*SNR*) ratio to be maximized is given by:

$$SNR = 10\log_{10}\frac{source\quad power}{quantization\quad error} = 10\log_{10}\frac{\sigma^2_{source}}{mse} \tag{5}$$

The rest of this paper is organized as follows. In section 2, definitions and properties of LSP parameters are presented. In section 3, the SVQ method used for the quantization of LSP coefficients is detailed. The SA algorithm is presented in section 4. Simulation results and discussions are given in section 5. Section 5 is dedicated to the conclusion.

2 LSP Properties

The linear predictive coding (LPC) method [38] is one of the most popular approaches for describing the time varying short-term spectrum of the speech signal. In many speech coding systems, LPC coefficients are transformed to the Line Spectrum Pairs (LSP) parameters [32] which are very effective representation for quantization of the LPC information. These parameters are preferable because they have a relatively low spectral sensitivity. This can be attributed to the intimate relationship between the LSP's and the formant frequencies. Accordingly, LSP's can be quantized taking into account spectral features known to be important in perceiving speech signals. In addition, LSP's lend themselves to frame-to-frame interpolation with smooth spectral changes because of their frequency domain interpretation. The LSP are related to the poles of the LPC filter (or the zeros of the inverse filter) in the Z-plane. For a p th-order LPC analysis, the Z-transform of the LPC inverse filter is denoted by:

$$A_p(z) = 1 + a_1 z^{-1} + \ldots + a_p z^{-p} \tag{6}$$

The parameters $\{a_i\}$ $i = 1,2,\ldots,p$, are commonly referred to as the LPC coefficients [38],

From (1) two new polynomials are defined:

$$\left.\begin{matrix} P(z) \\ Q(z) \end{matrix}\right\} = A_p(z) \pm z^{-(p+1)} A_p\left(z^{-1}\right) \tag{7}$$

The roots of these polynomials are usually called the Line Spectrum Pairs (LSP). These polynomials have the following properties:

All zeros of LSP polynomials are on the unit circle.

Zeros of P(z) and Q(z) are interlaced with each other on the unit circle.

The minimum phase property of $A_p(z)$ can be easily preserved if the first two properties are intact after quantization.

Some important properties are described in detail in [32].

The *10*th-order linear prediction corresponds to the frequency range of narrowband speech coders [39, 40].

3 Split vector quantization of LSP parameters

In this section we will present the SVQ definitions used for LSP coefficients quantization.

An m part, n-dimensional SVQ [2] operating at b bits/vector, divides the vector space, R^n, into m lower dimensional subspaces, $\{R_i^{n_i}\}_{i=1}^m$, where $n = \sum_{i=1}^m n_i$. Independent codebooks, $\{C_i\}_{i=1}^m$, operating at $\{b_i\}_{i=1}^m$ bits/vector, where $b = \sum_{i=1}^m b_i$, are then designed for each subspace. In order to quantize a vector of dimension n, the vector is split into subvectors of smaller dimension. Each of these subvectors is then encoded using their respective codebooks. The memory and computational requirements of the SVQ codebook are smaller than that of an unstructured VQ codebook. In terms of the number of floating point values for representing the SVQ codebooks as opposed to that of unstructured VQ:

$$\sum_{i=1}^{m} n_i 2^{b_i} \le n 2^b \tag{8}$$

while the effective number of code-vectors of the resulting product codebook is the same as that of unstructured VQ at the same bitrate:

$$\prod_{i=1}^{m} 2^{b_i} = 2^{b} \tag{9}$$

Therefore, the computational complexity and memory requirements for SVQ can be reduced considerably by splitting vectors into more parts.

In our study, the LSP parameters vector of dimension 10 is split into three sub-vectors, with the first sub-vector containing the three lowest LSP's, the second sub-vector containing the three middle LSP's and the final sub-vector containing the four highest LSP's [28].

4 Simulated Annealing

This section introduces the principle of SA algorithm suitable for solving the problem of VQ codebook design..

Simulated annealing is the computer modeling of the annealing process. By appropriately defining an effective temperature for the multivariable system, simulated annealing can solve a wide collection of optimization problems. Kirkpatrick et al. [12] were the first to use simulated annealing to solve such optimization problems. Starting from an initial state and with an initial temperature T_0, the simulated annealing proceeds as follows: Alter the state by a random perturbation, and compute the resulting change in the cost function, ΔE. If $\Delta E \le 0$, then the perturbed state is accepted as the new state. If $\Delta E > 0$, then the perturbation is accepted with probability $p(\Delta E) = \exp(-\Delta E / T)$. The state of the system is repeatedly perturbed until either a fixed number of attempts are made or a minimum number of attempts are accepted. The temperature T is then reduced to the next lower temperature, and perturbations are again carried out. The number of perturbations attempted at each temperature and the sequence of temperatures is called the annealing schedule. Kirkpatrick et al. [41] recommended that the annealing schedule be developed by trial and error for a given problem, and chose $T_n = (0.9)^n T_0$, for the VLSI partitioning problem. Hajek [41] recommended the schedule $T_n = C/\log(n+1)$ since this helps guarantee the global minimum. The primary advantage of simulated annealing is its ability to avoid local minima in its search for the state with globally minimum energy. Changes that both decrease and increase the cost function are accepted, making escape from local minima possible. In the next section codebooks design based SA will be tested.

5 Simulation results and discussions

A key issue in VQ is the design of the codebook. Usually, VQ codebooks designed using GLA [9] and LBG [10] algorithms, as stated in the introduction, can get trapped in local minima. Here we will investigate the use of SA method to optimize codebooks for SVQ (3-3-4) with LSP parameters and compare its performance with GLA and LBG in terms of *mse*, *SNR*, number of iterations and time execution.

The LSP coefficients where generated from the ITU-T G.729 standard, which operates at 8 kbits/s[41]. The speech used is extracted from the TIMIT database [42]. The total number of vectors used for the training sequence is 229829.

Here, vectors of dimension 10 representing speech LSP parameters are splitted into three subvectors of dimensions 3, 3 and 4 respectively. The bit allocation for each subvector is 8 bits with a total number of 24 bits/vector.

The annealing schedule used is the common $T_n = (0.9)^n T_0$, with T_0 selected to achieve 99% acceptance. Twenty five acceptances or rejections were required before decreasing the temperature.

Table 1 summarizes the results obtained with the GLA algorithm. It shows the $mse_{initial}$ and mse_{final} when the mse (eq. 4) is used as a cost function. Four cases are considered for each sub-vector of dimension n. Results obtained, when SNR (eq. 5) is considered as cost function to be maximized, are also reported in Table 1. The $SNR_{initial}$ and SNR_{final} are given considering four cases as mse is used cost function.

The total number of iteration and time execution is also given for each case.

The initial codebooks, for the three sub-vectors considered here, are generated randomly from the training sequence. Codebook index corresponds to the initial codebook and can be any integer value.

Tests for the LBG algorithm, reported in Table 2, are given by the statistical properties corresponding to each subvector represented by means and variances and are. Table 2 shows also the optimal *mse* and *SNR* and the corresponding number of iterations and time execution for each subvector of dimension n and size K=256.

The results obtained for SA algorithm are summarized in Table 3. The initial temperature T_i and its corresponding mse_{initia} are given for each subvector of dimension n when the mse is used as a cost function. The same thing for the final T_f and its corresponding mse_{final} when the SNR is used as a cost function. It is also given in the same table the corresponding number of iterations and time execution for each vector of dimension n.

Table 1. Results obtained with GLA for codebooks of dimension n and size K=256 of speech LSPs

Code-book	$mse_{initial}$	mse_{final}	$SNR_{initial}$	SNR_{final}	Codebook Index.	iterations	Time
	0.001706	0.000157	12.734338	23.097418	0	14	1.152 s
$n=3$	0.000130	0.000073	23.921389	26.416121	50	12	0.981 s
K=256	0.000214	0.000062	21.748373	27.116219	100	9	0.741 s
	0.011953	0.000829	4.280092	15.866799	400	23	1.883 s
	0.002855	0.000418	16.250179	24.590706	0	24	1.983 s
$n=3$	0.000590	0.000299	23.095110	26.054218	50	9	0.742 s
K=256	0.000810	0.000232	21.722563	27.145992	100	9	0.741 s
	0.010768	0.001036	10.484547	20.650553	400	13	1.072 s
	0.002654	0.000377	17.164139	25.634670	0	27	2.213 s
$n=4$	0.000799	0.000355	22.378906	25.897436	50	15	1.241 s
K=256	0.000953	0.000448	21.614010	24.888277	200	7	0.58 s
	0.005073	0.000689	14.350595	23.024261	300	15	1.231 s

Table 2. Results obtained with LBG for codebooks of dimension n and size K=256 of speech LSPs

Codebook	Mean	Variance	mse_{final}	SNR_{final}	iterations	Time
$n=3$ K=256	0.509195	0.032024	0.000047	28.294531	92	2.233 s
$n=3$ K=256	1.288222	0.120389	0.000194	27.930449	90	1.922 s
$n=4$ K=256	2.304135	0.138155	0.000222	27.936451	86	1.672 s

Table 3. Results obtained with SA for codebooks of dimension n and size K=256 of speech LSPs

Code-book	T_i	T_f	$mse_{initial}$	mse_{final}	$SNR_{initial}$	SNR_{final}	iterations	Time
$n=3$ K=256	1	$9.89\ 10^{-36}$	0.011953	0.000034	4.280031	29.720501	19819	14min38s
$n=3$ K=256	10	$9.74\ 10^{-36}$	0.010768	0.000136	10.484393	29.469749	20299	15min26s
$n=4$ K=256	500	$9.88\ 10^{-36}$	0.005534	0.000186	13.972967	28.714369	21238	14min55s

Comparing Tables 2 and 3, we can see that SA algorithm gives better results than LBG in terms of mse_{final} for the three codebooks considered. We can also notice that the SA algorithm is more CPU time consuming compared to the LBG algorithm.
Comparing Tables 1, 2 and 3, the results obtained with GLA in terms of mse_{final}, SNR_{final} are worse than the other two methods, they are highly dependent on the initial codebook chosen.

Comparing Tables 2 and 3, we notice that he LBG algorithm is faster but less efficient than SA. This was expected because a descent method (LBG in our case) is theoretically less time consuming than Sa but less efficient. A descent method is often trapped in a local minimum especially when the objective function to minimize in our case, the *mse*, has several minima; this is due to the search criterion of a descent method. It evolves in its search for the solution (quantization codebook in our case) through the optimal set of solutions by not accepting a lower cost solution than the current solution from one step to another; it stops the search if a minimum is met but not necessarily the global minimum.

The performance of a descent method is directly related to the quality of the initial solution from which begins the search procedure for the optimal solution, that's what we found for GLA. GLA is a descent method, which is identical to LBG (same optimality conditions), but the major problem in the GLA algorithm is the choice of the initial codebook. We chose to create the initial codebook of GLA randomly and we noticed that some GLA codebooks approached the initial results obtained by LBG and SA, but more often in practice it is not easy to find an initial codebook that ensure the convergence of GLA to an optimal codebook, it may be more difficult than the original problem and therefore a waste of time and more without reaching suitable results.

SA performs better than LBG and GLA; this is due to the global search of SA. It accepts solutions that improve the cost of the objective function (in our case *mse* or

SNR) and also in a controlled manner (probabilistic) solution which degrade it. The performance of SA is directly related to the cooling scheme selected and the number of iterations per temperature. For a given time, the simulated annealing will approach as possible the optimal solution. In some problems the time required for simulated annealing performance could be seen as a disadvantage, but for the quantization problem this is not a waste of time because the quantization codebook is designed eternally.

6 Conclusion

Simulated annealing is a powerful optimization procedure that achieves near globally-minimum-cost solutions to many optimization problems. In this paper, we attempted to apply SA to improve the quality of codebooks SVQ for the quantization of spectral parameters represented by LSPs. SA provided the best SNR and mse results, avoiding the initial codebook dependence found when using the GLA.

Simulated annealing, while itself, too time consuming, does serve to obtain a near globally-optimum solution for codebook design. Future research will focus on more sophisticated algorithms based genetic algorithms and Tabu search.

References

1. Gersho, A., Gray, R. M.: Vector Quantization and Signal Compression. Kluwer Academic Publishers (1992)
2. Gray, R. M.: Vector quantization. Mag. IEEE Acou. Spee. Sig. Pro. 1, 4-29 (1984)
3. Nasrabadi, N. M., King, R. A.: Image coding using vector quantization:A review. IEEE. Tran. Com. 36, 957-971(1988)
4. R.M. Gray, D.L. Neuhoff, Quantization, IEEE Trans. Inform. Theory 44 (6) (1998) 2325–2383.
5. K. Sayood, Introduction to Data Compression, Morgan Kaufmann Publishers, San Francisco, 1996.
6. Mahalanobis, P.C. :On the generalized distance in statistics, in: Proc. Indian Nat. Inst. Sci. Calcutta, 2, 49–55 (1936).
7. Itakura, F. :Minimum prediction residual principle applied to speech recognition, IEEE Trans. Acoust. Speech Signal Process. ASSP-23 (1) (1975) 67–72.
8. Makhoul, J. Roucos, S. Gish, : Vector quantization in speech coding, Proc. IEEE 73 (1985) 1551–1588.
9. Lloyd, S.P.: Least square quantization in PCM. IEEE Trans. Inform. Theo.28, 129–137. (1982)
10. Linde, Y., Buzo, A., Gray, R.M.: An algorithm for vector quantizer design. IEEE Trans. Commun. COM. 28, 84–95 (1980)
11. Lloyd, S.P. :Least square quantization in PCM, IEEE Trans. Inform. Theory IT-28 (2) 129–137 (1982)
12. S. Kirkpatrick, C. D. Gellatt, Jr., and M. P. Vecchi. “Optimization by simulated annealing,” Science. vol. 220. 671 -680 (1983).
13. Vaisey J. and Gersho, A. :Simulated annealing and codebook design,” in Proc. IEEE ICASSP .,1176-1179 (1988)

14. Flanagan, J. K.. Morrell, D.R. Frost. R. L. Read, C. J and Nelson, B. E. V:ector quantization codebook generation using simulated annealing, in Proc. IEEE ICASSP, 3, 1759- 1762 (1989)
15. Nasrabadi, N. M., Feng, Y. "Vector Quantization of Images Based Upon the Kohonen Self-organizing Feature Maps." Proceedings IEEE International Conference on Neural Networks, 1, 101-108 (1988)
16. Wu, F. H., Ganesan, K. "Comparative Study of Algorithms for VQ Design Using Conventional and Neural-net Based Approaches." Proceedings International Conference on Acoustics, Speech, and Signal Processing (ICASSP'89), 2, 751-754(1989)
17. Madeiro, F., Vajapeyam, M. S., Morais, M. R., Aguiar Neto, B. G., and Alencar, M. S. "Multiresolution Codebook Design for Wavelet/VQ Image Coding". Proceedings of the 15th International Conference on Pattern Recognition (ICPR'20000 Barcelona, 3, 79-82, (2000).
18. Chang, C.-H., Xu, P., Xiao, R., Srikanthan, T. New Adaptive Color Quantization Method Based on Self-Organizing Maps. IEEE Transactions on Neural Networks, 16 (1), 237-249, (2005).
19. D.E. Goldberg, "Genetic Algorithms in Search, Optimization and Machine Learning," Addison-Wesley, Reading, 1989.
20. V. Delport, N. Koschorreck, "Genetic Algorithm for Codebook Design in Vector Quantization", Electronics Letters, 31(2), 84-85 (1995)
21. Pan, J. S. Mcinnes, F. R.. Jack, M. A :VQ Codebook Design Using Genetic Algorithms", Electronics Letters, 31(17), 1418-1419 (1995)
22. Hsiang-Cheh Huang, Jeng-Shyang Pan, Zhe-Ming Lu, Sheng-He Sun, Hsueh-Ming Hang. Vector quantization based on genetic simulated annealing. Signal Processing, 81, 1513-1523 (2001)
23. Yuan, Y J Zhou, Q Zhao. P H :Vector quantization codebook design method for speech recognition based on genetic algorithm, .Proceedings of the 2010 2nd International Conference on Information Engineering and Computer Science. Wuhan, 1-4 (2010)
24. Santo, P.H.E.; Albuquerque, R.C.; Cunha, D.C.; Madeiro, F.;, On Frequency Sensitive Competitive Learning for VQ Codebook Design, Neural Networks, 2008. SBRN '08. 10th Brazilian Symposium on , 135-140, 26-30 (2008)
25. Kang, S.. Shin, Y and Fischer, T. R. Low-complexity predictive trellis coded quantization of speech line spectral frequencies, IEEE Trans. Signal Processing, 52, 2070-2079 (2004)
26. So S.and Paliwal, K.K.: A comparative study of LPC parameter representations and quantization schemes for wideband speech coding", Digital Signal Processing, 17(1), 114-137(2007)
27. Wakita, H. :Linear prediction voice synthesizers: Line spectrum pairs (LSP) is the newest of several techniques, Speech Technol., (1981).
28. Paliwal, K.K. Atal, B.S. :Efficient vector quantization of LPC parameters at 24 bits/frame, IEEE Trans. Speech Audio Process. 1 (1) 3–14 (1993)
29. A.M. Smith, J.P. Ashley, M.A. Jasiuk, W. Peng, Normalization and polygon error detection for split VQ of line spectral frequencies, in: IEEE Speech Coding Workshop, 125–27 (2000)
30. Nordén, F. Eriksson, T: .On split quantization of LSF parameters, in: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Montreal, I-157–I-160 (2004)
31. Markel, J. D. and Gray, A. H. Jr.: Linear Prediction of speech. New York: Springer-Verlag, 1976
32. Itakura, F. :Line spectrum representation of linear predictor coefficients of speech signal's," J. Acoust. SOC. Amer . , 57, S35(A), (1975)
33. Kang G. S . and. Fransen, L. J :Low bit rate speech encoders based on line spectrum frequencies (LSFs)," Naval Res. Lab., Rep. 8857, (1984)
34. Gray R. and Neuhoff, D. :Quantization, IEEE Trans. Inform. Theory, 44, 2325–2383(1998)

35. Paliwal, K.K. Atal, B.S. :Efficient vector quantization of LPC parameters at 24 bits/frame, in: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, 1991, pp. 661–664.
36. 3rd generation partnership project; Technical specification group services and system aspects; Mandatory speech codec speech processing functions; Adaptive multi-rate (AMR) speech co-dec; Transcoding functions (Release 5), Technical Specification TS 26.090, 3rd Generation Partnership Project (3GPP), June 2002.
37. Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms, Tech. Rep. Standard ES 201 108, European Telecommunications Standards Institute (ETSI), April 2000.
38. Makhoul, J.: Linear prediction: A tutorial review speech." Proc. IEEE. 63,124-143(1975)
39. ITU, ITU-T G.723.1: Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s, ITU 1996.
40. ITU, ITU-T G.729: CS-ACELP Speech Coding at 8 kbit/s, ITU 1998
41. Hajek, B.: A tutorial survey of theory and applications of simulated annealing," in Proc. 24th Conf. on Decision and Control, 755-760 (1985)
42. NIST,Timit Speech Corpus, NIST (1990)

Pulsar Synthesis Revisited: Considerations for a MIDI Controlled Synthesiser

Thomas Wilmering[1], Thomas Rehaag[2], and André Dupke[3]

[1] Centre for Digital Music (C4DM)
Queen Mary University of London
London E1 4NS, UK
thomas.wilmering@eecs.qmul.ac.uk

[2] Intelligent Sounds & Music
Cologne 51065, Germany

[3] Hamburg-Audio
Periana 29710, Spain

Abstract. In this paper we present an implementation of a software synthesiser based on pulsar synthesis to be used in conventional digital audio workstations supporting common plugin standards. After reviewing basic pulsar synthesis, we describe limitations of this synthesis technique and novel parameters we developed to overcome these for the design of a MIDI controlled implementation. The developed keyboard instrument can be easily played by composers and music producers familiar with software synthesisers using traditional synthesis techniques based on virtual oscillators. We also discuss aesthetic considerations in the design, the spectra of complex grain waveforms, and the effect of parameter changes on pulsar train spectra.

Keywords: sound synthesis, pulsar synthesis, granular synthesis

1 Introduction

Sound synthesis and manipulation based on granular synthesis (GS) has been investigated in great detail since its first computer-based implementation by Curtis Roads in 1978 [1]. Newer implementations range from sound generators for grain clouds to digital audio effects based on sound granulation [2]. An extensive overview of granular synthesis techniques is given in [3]. However, although granulation and granular synthesis is a widespread technique in contemporary electronic music composition, synthesisers based on pulsar synthesis (PS) are less common. This is especially the case for implementations that can be easily integrated into digital audio workstations as plug-ins to be used in the same fashion as virtual analog synthesisers

PS is a type of granular synthesis whose name originates from spinning neutron stars that emit periodic signals in the range of 0.24 to 642 Hz [4]. Basic

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

pulsar synthesis generates a periodic pulsar train controlled by various parameters which we describe in more detail in section 3. The aim of our research is to extend PS, and to overcome the limitations inherent in the technique when implemented as a keyboard instrument where the pitch is controlled by MIDI note numbers. The goal is not only the creation of a composition tool for experimental sound design, but also to make PS accessible to, and intuitively usable for music producers of popular electronic music, stressing some of its apparent resemblances to virtual analog synthesis.

After briefly describing granular synthesis and its musical application and reviewing basic pulsar synthesis, we describe the extended PS and the aesthetic considerations taken into account during the development of a plug-in synthesiser in conjunction with *hamburg-audio*[4]. Lastly we discuss the sonic capabilities by means of complex grain waveform spectra, the effect of parameter changes on pulsar train spectra, and sequencing in the microsound domain. Audio examples accompanying this paper can be found online[5].

2 Granular Synthesis

GS is based on the theory of acoustical quanta by British physicist Dennis Gabor, who suggested that every sound can be decomposed to a family of functions derived from time and frequency shifts of a single Gaussian particle. Gabor developed a mathematical representation for acoustical quanta by relating a time-domain signal with a frequency-domain spectrum [5][6]. The duration of a grain of sound is usually in the range of 1 to 100ms, ranging near the threshold of human perception. Furthermore, a grain is characterised by its waveform w shaped by an envelope v.

The combination of large numbers of grains over time makes it possible to create sound patterns and *grain clouds* resulting in atmospheric sounds [3]. The basic form of a grain generator is shown in figure 1.

The first person to use Gabors theory as a composition tool was Xenakis [7]. Musical pieces to mention are for example Metastasis (1954), Concret PH (1958) and Analogique A-B (1959), the latter being described in [8]. Curtis Roads did further research in the field of GS and created various related compositions and computer programs. He developed software implementations such as *Cloud Generator* and *Pulsar Generator* [3][4]. Other important composers in this context are Paul Lansky and Barry Truax. Until the end of the 20th century, GS could mostly be found in the works of composers linked to scientific research institutions and in compositions outside of the world of popular music. However, new genres and subcultures emerged since the beginning of this century with composers having different degrees of knowledge about the institutional framework of computer music. Inspired by the works of the established composers, music styles such as *glitch* or *IDM* (Intelligent Dance Music) are heavily influ-

[4] http://hamburg-audio.com
[5] http://isophonics.net/content/pulsar-synthesis

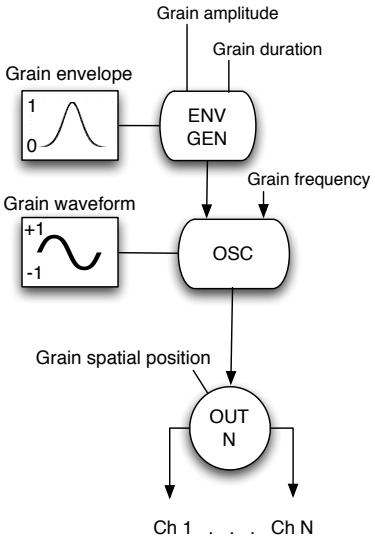


Fig. 1. Basic grain generator consisting of a Gaussian grain envelope generator and a sinusoidal grain waveform. The grains can be spatially placed in N channels.

enced by GS and *Granulation*. This development has also been discussed in the musicological literature [9][10][7].

In recent years granular synthesis and granulation-based effects have found their way into popular music and music production tools with a variety of digital audio effects inspired by the aforementioned newly emerging genres and subgenres of electronic music. However, PS has as of today gained less attention since its first appearance.

3 Pulsar Synthesis

PS has first been presented as a computer-based granular synthesis technique by Curtis Roads and Alberto de Campo in 1999 [4]. The pulses and pitched tones produced with PS are similar to those of earlier analog musical instruments, e.g. the Ondioline or the Hohner Elektronium which are based on filtered pulse trains [11][4]. Nevertheless, PS is implemented in the digital domain, therefore taking advantage of the processing power and flexibility of modern computer systems.

In its simplest form a pulsar train (as shown in figure 2) is controlled by two main parameters, the fundamental frequency (pulsar frequency):

$$f_p = \frac{1}{p} \tag{1}$$

$$p = d + s \tag{2}$$

where the period p of the pulsar consists of the *pulsaret* width (duty cycle) d, and the intergrain time s.

The duty cycle frequency (formant frequency) is described by:

$$f_d = \frac{1}{d} \tag{3}$$

which determines the width of the pulsaret within the pulsar.

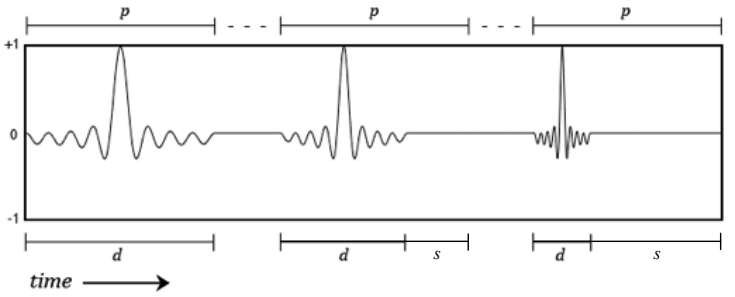


Fig. 2. A pulsar consists of a *pulsaret* of width d and a following silent part (intergrain time s). The period p of the fundamental determining the pitch is an independent parameter from the pulsaret width.

The periodic pulsar train G can be expressed as the convolution of the pulsaret with an impulse train:

$$G_{w,v,d,p}(t) = w_d v_d * \sum_{k=-\infty}^{\infty} \delta(t - kp) \tag{4}$$

where w_d and v_d are scaled versions of the pulsaret waveform and envelope with length d, and δ is the Dirac delta function.

As opposed to pulse width modulation (PWM) in analog synthesisers, where the duty cycle of a rectangular waveform is set by a ratio to the fundamental period, in PS the duty cycle is an independent parameter from the fundamental frequency. Moreover, the pulsaret is characterised by the pulsaret waveform w and the pulsaret envelope v. The pulsaret waveforms and envelopes can be of arbitrary shape. However, Roads [4] proposed some standard waveshapes with the initial introduction of PS and investigated the effect of the grain envelope on the grain's spectrum. The standard waveforms include *sine* and *multicycle sine*, as well as bandlimited pulses and cosmic pulsar waveforms stressing the synthesis technique's relationship to waveforms emitted by neutron stars. Typical envelope shapes are for example Gaussian, sine, linear or exponential decay or attack. The envelope causes a resonant main band and several sidebands, smearing the original waveform spectrum [12]. Figure 2 shows pulsars of constant pitch with varying duty cycle frequency; the pulsaret waveform is a band-limited pulse.

As f_p and f_d are independently variable parameters, we may encounter the case of $d > p$. In this case we can apply overlapped pulsaret-width modulation

(OPulWM), where several grains (pulsarets) overlap. This overlap is defined as the time interval during which two or more grains are played simultaneously, and can be calculated by the difference between the grain rate and grain duration [13]. A different approach to deal with this problem is to cut off the pulsar and spawn a new one without overlapping. However, in practice both approaches have disadvantages. While OPulWM generally leads to cancellation at higher numbers of overlapped grains (and increased CPU load), cutting of a grain may lead to sudden changes in the amplitude, introducing unwanted high frequencies into the spectrum (this effect may be dampened by a cross-fade parameter at the cutoff point). In section 4.6 we describe a hybrid synthesis approach used in our implementation which presents an optional compromise between true PS an playability throughout all octaves.

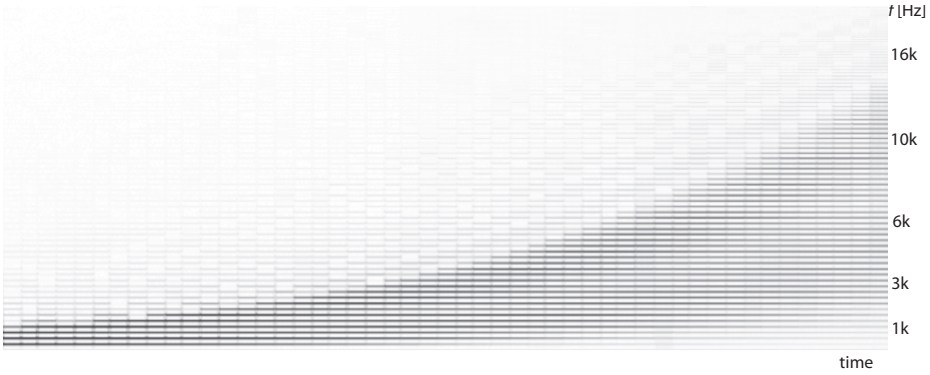


Fig. 3. Spectral effect of increasing the duty cycle frequency at a constant fundamental frequency (pulsaret width modulation).

4 Pulsar Synthesis for a MIDI Controlled Synthesiser

For the design of a PS synthesiser that can be played as a keyboard instrument we introduced a number of novel parameters, some to improve the playability, others purely for aesthetic considerations. In this section we present some of these parameters, the motivation behind their introduction, and the sonic implications. We implemented the synthesis in the *Nuklear* synthesiser as an *Audio Unit*[6]/*VST*[7] plug-in with four parallel pulsar train generators. Its graphical user interface (GUI) is shown in figure 4.

In addtion to the sound generators the synthesiser features various modulation sources which can be mapped to arbitrary parameters. These include 8

[6] https://developer.apple.com/library/mac/#documentation/MusicAudio/Conceptual/AudioUnitProgrammingGuide/Introduction/Introduction.html

[7] http://ygrabit.steinberg.de/~ygrabit/public_html/index.html

Fig. 4. Graphical user interface of *Nuklear*, a plug-in synthesiser based on Pulsar Synthesis

ADSHR envelopes, 8 low frequency oscillators (LFO) and control sequences that can be programmed in a 16-step sequencer. Moreover, the sum on the output can be shaped by two filters and an effect section consisting of a delay effect and a distortion effect. In this paper, however, we focus specifically on the pulsar train generators and the parameters we introduced to extend PS to our needs.

4.1 Pulsaret Waveforms

In addition to the standard PS waveforms (see section 3) we implemented several other waveforms known from classic virtual oscillators, such as *sawtooth*, *rectangular* and *triangular*. Furthermore, we added experimental waveforms, among them a selection of shapes based on wavelet functions. It should be noted that in our case we perform a sonification of wavelet functions rather than mathematical operations related to their original purpose, the wavelet transform for multi-resolution signal analysis. We chose the wavelet shapes from an aesthetic point of view and achieved some interesting results with regards to composition and sound design. In our experiments we investigated acoustic characteristics of compactly supported orthonormal wavelets as introduced by Daubechies [14], and biorthogonal wavelets, a family of wavelets with the property of linear phase [15]. The wavelet shapes that are included in our implementation were selected to cover a wide range of sounds. Figure 5 shows some of the waveforms and their respective spectra.

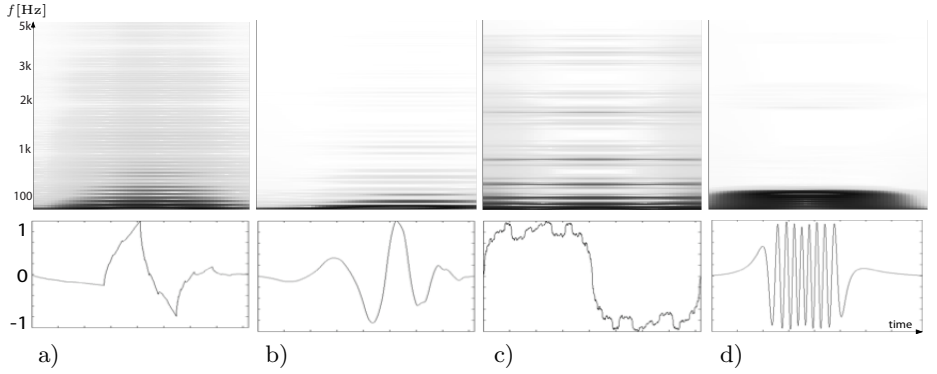


Fig. 5. Pulsaret waveforms with corresponding spectra above. a) based on Daubechies wavelet 2; b) based on Daubechies wavelet 5; c) based on biorthogonal decomposition wavelet 3.5; d) cosmic gravitational wave. The spectrograms are 2048-point fast Fourier transform plots with a Blackmann-Harris window. The duty cycle frequency is 10.87Hz.

To shape the waveforms we implemented the envelope shapes shown in figure 6 in addition to a rectangular envelope with a constant value of 1.

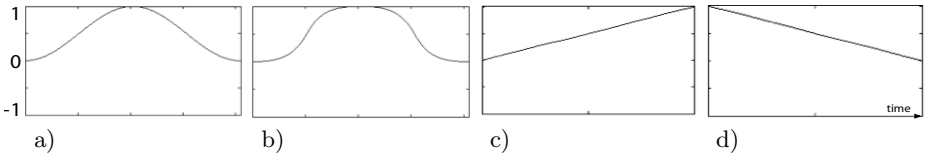


Fig. 6. Pulsaret envelopes in *Nuklear*: a) Hann type I; b) Hann type II; c) linear attack; d) linear decay.

4.2 Stereo Width

A stereo widening effect can be set independently for each of the four parallel pulsar trains. The effect is achieved by alternating the pulsars between the stereo channels (see figure 7). This is implemented as a parameter allowing intermediate settings. Low settings of this parameter only decrease the amplitude in an alternating fashion between the channels, instead of muting every other pulsar completely.

At low fundamental frequencies in the infrasonic range the alternating pulsars are clearly distinguishable, while at higher notes the stereo widening effect is perceived. Moreover, even at low settings harmonics of half the fundamental frequency are introduced to the spectrum. This is due to the fact that effectively an amplitude modulation at half the fundamental frequency is performed with a 180° phase shift between the stereo channels. At width settings above the middle position the alternating pulses are inverted and mixed with the opposite channel.

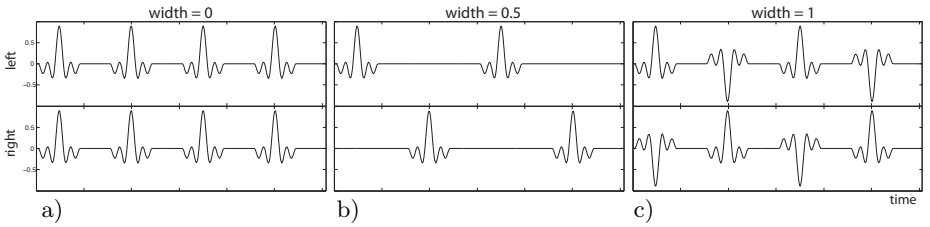


Fig. 7. Stereo width parameter: At 0 every pulsar plays on both channels (a). At the centre position the pulsars alternate between the channels (b). At 1 the alternating pulses are mirrored negatively onto the other channel (c). This parameter also allows intermediate settings.

4.3 Pitch-Dependent Pulsar Phase

Another novel parameter we introduce is the *pitch-dependent pulsar phase*. We define this parameter as a time shift of every other pulsar within the fundamental period in the range of 0° to 360°. Figure 8 shows the effect this time-shift has on the pulsar train, both in the time and frequency domain. A phase shift of

360° translates to every second pulsar coinciding with the next pulsar, effectively transposing the pulsar train down by one octave.

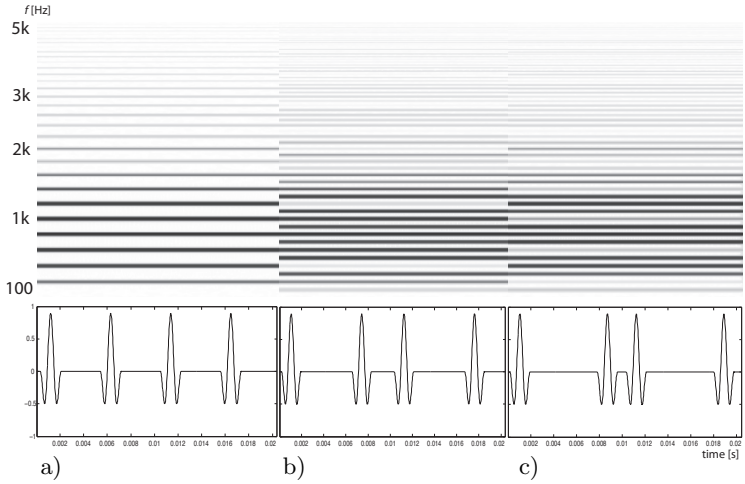


Fig. 8. Effect of the pitch-dependent pulsar phase shift on the pulsar train. The spectra above are produced by playing continuous loops of the pulsar sequences below. a) phase = 0°; b) phase = 90°; c) phase = 180°.

This phase shift technique introduces harmonics at half the original fundamental frequency, with varying magnitudes for the partials depending on the phase-shift amount. Modulating the parameter with an LFO or envelope curve results in an effect reminiscent of a classic *phaser*. However, the common *phaser* effect consists of a time-modulated additive delay line in the range of up to 2ms, independently from the note's pitch [16]. In the infrasonic domain the pulsar phase parameter produces rhythmic changes in the pulsar train.

4.4 Grain Overlap

As mentioned in section 3, when implementing a granular synthesiser one needs to take into account the case of overlapping grains. In the case of PS this is relevant when the fundamental frequency f_p exceeds the duty cycle frequency f_d. Overlapping grains in PS can produce a smooth sound due to the negative intergrain time, while they at the same time largely preserve the overall formant structure of the pulsar train. However, high numbers of overlapping grains may lead to cancellation and high CPU load.

In *Nuklear* it is possible to define an *overlap limit*. When this limit is reached, the pulsarets are scaled in such a way, that a set maximum number of pulsarets fits in to the fundamental cycle. While this technique changes the sound characteristics of true PS, it allows the user to play the synthesiser over all octaves without having to worry about undesired missing notes in the higher registers.

Figure 9 shows a pulsar train with pulsarets scaled to fit into the fundamental period p with an overlap limit of 0 (a), and overlapping pulsarets with unchanged duty cycle frequency f_d (a).

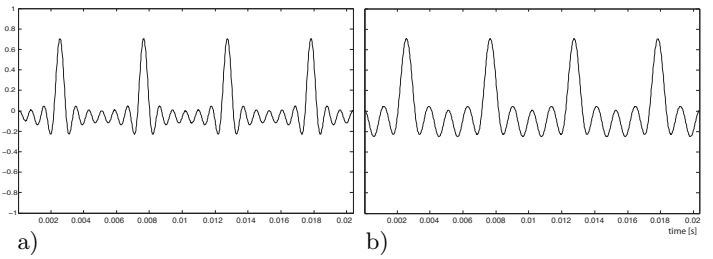


Fig. 9. Pulsar train with a pulsaret waveform of a bandlimited pulse with 6 harmonics. a) overlap limit = 0, the pulsaret is scaled to fit into the fundamental period ($d = p$); b) the pulsarets overlap and keep the original duty cycle frequency f_d.

Figure 10 shows how allowing grains to overlap preserves the formant structure (a). Automatic scaling of the duty cycle period to fit into the fundamental period to avoid grain overlap results in different spectra. In the latter case the pulsar train generator behaves similarly to a virtual oscillator using the pulsaret waveform (in this example d is scaled to $0.95p$).

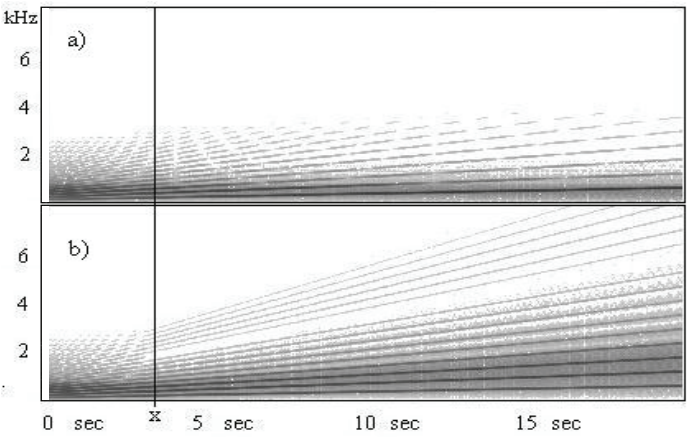


Fig. 10. A pulsar train generated with a prototype implementation. f_p rises linearly from 100 Hz to 600 Hz. The duty cycle frequency f_d is at 200 Hz, the pulsarets are two sine wave cycles shaped by a Gaussian envelope. At the point marked x, f_p is $0.95 f_d$. a) the grains start to overlap, b) the pulsarets are scaled to $d = 0.95p$ to avoid overlap.

4.5 Sequencing in the Microsound Domain

Pulsar masking is the controlled deletion of pulsars from the pulsar train replacing it with silence. Roads [4] proposes three forms for this technique: *burst*, *channel*, and *stochastic masking*.

Burst masking mutes pulsars at regular intervals at a given *burst ratio b:r*, where b defines the burst length and r is the rest length in pulsaret periods. For instance, a *burst ratio* of 3:1 produces a sequence of 3 pulsarets and one period of silence, which can be denoted by a binary sequence 111011101110, etc. (Figure 11b). The effect is a form of amplitude modulation, where the fundamental frequency is broken up by a subharmonic factor $b+r$. If the fundamental frequency is in the infrasonic range, rhythmic patterns are perceived. *Channel masking* is defined as the muting of pulsarets on two channels in an opposite manner. Therefore, the stereo width parameter at its middle setting as described in section 4.2 can be seen as channel masking with a *burst ratio* of 1:1. *Stochastic masking* mutes pulsars randomly according to a given ratio, i.e. the ratio of the number of outcomes of 1 (play) to the number of outcomes of 0 (mute), when all outcomes are regarded as equally likely.

Building on the idea of *burst masking* we implemented a binary pulsar masking sequencer we call *microsequencer*. In its current form it allows a masking sequence of up to 8 pulsar cycles in which pulsars can be set to either *on* or *off*. Figure 11c) shows the pulsar pattern 10111001, and the resulting spectrum of a note played with this masking setting.

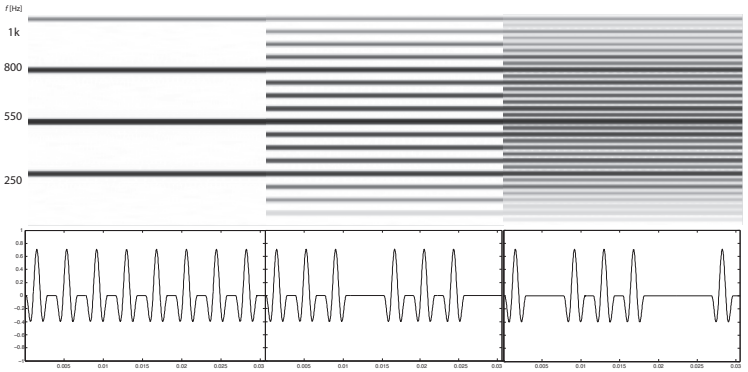


Fig. 11. Spectral effect of the *microsequencer*. The spectra above are produced by playing continuous loops of the pulsar sequences below. a) regular pulsar train without masking; b) masking every 4th pulsar introduces harmonics at $\frac{1}{4}$ of the fundamental frequency (sequence 1110); c) a more complex pulsar sequence (10111001).

4.6 Pulsar/Virtual-Analog Hybrid Synthesis

In addition to true PS our synthesiser offers a synthesis that can be described as a hybrid between PS and synthesis using *classic* virtual oscillators. This development is partly motivated by the problems with regards to notes with high fundamental frequencies and the resulting multiple grain overlap (see section 4.4). Moreover, due to the nature of PS many sounds lack low frequencies in the spectrum and the perceived *warmth* associated with them. We address this problem by lowering f_d relative to f_p, limiting the effect that low notes are not only perceived as a series of high-pitched clicks. By introducing a parameter h, we can gradually move between PS ($h = 1$) and virtual oscillator based synthesis ($h = 0$). Here, the duty cycle frequency f_d takes the form:

$$f_d(h, e, p) = hf_e + (1 - h)f_p \ , 0 \leq h \leq 1 \tag{5}$$

where h is the hybrid synthesis parameter setting and f_e is the duty cycle frequency setting on the GUI.

Note that if $h = 0$ then $f_d = f_p$, i.e. in this setting the pulsar train generator can be used in the same way as a virtual analog oscillator. Moreover, an oscillator may be defined as a special case of a pulsar train generator, where $f_p = f_d$. Thus, all the complex waveforms and parameters developed for PS as described in sections 4.1 to 4.5 are still available in this synthesis mode. Furthermore, by utilising multiple pulsar train generators and the *microsequencer* we can design unusual oscillators, for instance, by rotating through a sequence of waveforms. Figure 12 shows such a signal produced by employing all four pulsar train generators with $h = 0$, each of which use a different waveform and pulsar sequences (1000, 0100, 0010, 0001).

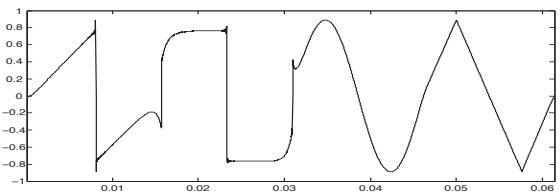


Fig. 12. Signal produced by employing four pulsar train generators each playing a different pulsar sequence with different waveforms. The microsequencer is set to a length of 4 and the sequences 1000, 0100, 0010, 0001. The duty cycle d is of the same length as the fundamental period p

5 Conclusions

We presented an implementation of a software synthesiser based on extended pulsar synthesis (PS). A number of novel parameters were introduced in order to

enhance the synthesis technique specifically for use in a keyboard instrument. We demonstrated that PS is well suited for the development of tools for music composition and production, not only within the academic musical framework, but also as an alternative to established virtual synthesisers used in music production studios. To overcome some of the subjective limitations of PS we proposed a hybrid synthesis technique which makes it possible to gradually move between PS and classic virtual oscillator based synthesis. Moreover, due to the nature of PS the synthesiser can be used to demonstrate the relationship of rhythm and pitch by comparing pulsar trains with fundamental frequencies in the infrasonic range with pulsar trains in the harmonic range. The former produces rhythmic patterns, the latter produces pitched notes.

The professional plug-in implementation *Nuklear* is based on this research and capable of producing a range of sounds different in character from established synthesisers. It received positive reviews in the professional literature, and has been awarded with the *Innovation Award* from *Computer Music* magazine [17].

References

1. C. Roads, "Granular synthesis of sound," *Computer Music Journal*, vol. 2, no. 2, pp. 61–62, 1978.
2. U. Zölzer (Ed.), *DAFX - Digital Audio Effects*, J. Wiley & Sons, second edition edition, 2011.
3. C. Roads, *Microsound*, MIT Press, Cambridge, MA, USA, 2001.
4. C. Roads, "Sound composition with pulsars," *Journal of the Audio Engineering Society*, vol. 49, no. 3, pp. 134–147, 2001.
5. D. Gabor, "Acoustical quanta and the theory of hearing," *Nature*, vol. 159, no. 4044, pp. 591–594, 1947.
6. M. J. Bastiaans and A. J. van Leest, "Gabor's signal expansion and the gabor transform based on a non-orthogonal sampling geometry," *6th International Symposium on Signal Processing and its Applications, Kuala Lumpur, Malaysia*, vol. 1, pp. 162–163, 2001.
7. P. Thomson, "Atoms and errors: Towards a history and aesthetics of microsound," *Organised Sound*, vol. 9, no. 2, pp. 207–218, 2004.
8. I. Xenakis, *Formalized Music (Revised Edition)*, Pendragon Press, Stuyvesant, NY, USA, 1992.
9. K. Cascone, "The aesthetics of failure: "post digital" tendencies in contemporary computer music," *Computer Music Journal*, vol. 24, no. 4, pp. 12–18, 2002.
10. P. Sherburne, "12k: Between two points," *Organised Sound*, vol. 9, no. 2, pp. 225–228, 2002.
11. L. Fourier, "Jean-Jacques Perrey and the ondioline," *Computer Music Journal*, vol. 18, no. 4, pp. 18–25, 1994.
12. D. Keller and C. Rolfe, "The corner effect," *Proceedings of the XIIth Colloquium of Musical Informatics, Gorizia, Italy*, 1998.
13. D. L. Jones and T. W. Parks, "Generation and combination of grains for music synthesis," *Computer Music Journal*, vol. 12, no. 2, pp. 27–33, 1988.
14. I. Daubechies, "Orthonormal bases of compactly supported wavelets," *Communications on Pure and Applied Mathematics*, vol. 41, no. 7, pp. 909–996, 1988.

15. A. Cohen, I. Daubechies, and J.-C. Feauveau, "Biorthogonal bases of compactly supported wavelets," *Communications on Pure and Applied Mathematics*, vol. 45, no. 5, pp. 485–560, 1992.
16. H. Speckmann, "Time-related sound processors," [Online]. Available: http://www9.dw-world.de/rtc/infotheque/sound_processors/soundprocessors.html, 2004.
17. *Computer Music*, Number 172. Future Publishing Ltd., 2011.

Knowledge Management On The Semantic Web: A Comparison of Neuro-Fuzzy and Multi-Layer Perceptron Methods For Automatic Music Tagging

Sefki Kolozali, Mathieu Barthet, and Mark Sandler

Centre for Digital Music
Queen Mary University of London
{sefki.kolozali,mathieu.barthet,mark.sandler}@eecs.qmul.ac.uk

Abstract. This paper presents the preliminary analyses towards the development of a formal method for generating autonomous, dynamic ontology systems in the context of web-based audio signals applications. In the music domain, social tags have become important components of database management, recommender systems, and song similarity engines. In this study, we map the audio similarity features from the Isophone database [25] to social tags collected from the Last.fm online music streaming service, by using neuro-fuzzy (NF) and multi-layer perceptron (MLP) neural networks. The algorithms were tested on a large-scale dataset (Isophone) including more than 40 000 songs from 10 different musical genres. The classification experiments were conducted for a large number of tags (32) related to genre, instrumentation, mood, geographic location, and time-period. The neuro-fuzzy approach increased the overall F-measure by 25 percentage points in comparison with the traditional MLP approach. This highlights the interest of neuro-fuzzy systems which have been rarely used in music information retrieval so far, whereas they have been interestingly applied to classification tasks in other domains such as image retrieval and affective computing.

1 Introduction

In the last decade, there has been extensive research on the development and use of the semantic web to make the web data interpretable, usable and accessible across a wide variety of domains. The key idea of this effort is to provide web content with conceptual background which is referred to as ontologies. For this purpose, the data models, such as ontology web language (OWL) and resource description format (RDF) have received considerable attention from researchers and the industrial sectors. Many research groups built ontologies manually to represent different types of data (e.g. music data, social data) within the formation of the semantic web [1]. Some examples of ontologies in the music domain are the music ontology[1] (MO) and the music performance ontology, grounded in the MO [22].

[1] http://musicontology.com/

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

The main disadvantage of the current ontology engineering process is that it cannot operate independently from human supervision. There is a growing interest for automated learning systems which can handle knowledge acquisition and also build ontologies from fast growing and large datasets [3], since current ontologies have an inflexible structure, and are incapable of handling these problems.

Social tags represent a potential high-volume source of descriptive metadata for music. Tags are useful text-based labels that encode semantic information about the music content (e.g. genres, instrumentations, geographic origins, emotions). In the music domain, popular web systems such as Last.fm[2] provide possibility for users to tag with free text labels an item of interest. Such metadata can either be used to train audio content-based classification systems for semantic annotation and retrieval, or likewise, automatic ontology generation. There has been recently a significant amount of research on content-based music similarity and tagging systems. Both fields use content-based descriptors extracted from audio signals. The Isophone dataset [25] provides an excellent opportunity to undertake reproducible research on large-scale music collection with readily-available mel-frequency cepstral coefficient (MFCC) features that can be jointly used with other datasets.

In this paper, we propose an audio tagging system based on neuro-fuzzy (NF) neural networks in comparison with the more traditional multi-layer perceptron (MLP) algorithm. The system was tested using the Isophone database in conjunction with Last.fm social tags. The use of neuro-fuzzy systems is driven here for further linking it with fuzzy spatial reasoning as an ontology generation solution. Hence we are motivated here by the comparison of the performance of NF networks relatively to another classifier, rather than by the obtention of state-of-the-art classification accuracies. Neuro-fuzzy systems have only been scarcely used in MIR (e.g. [29]) whereas they have shown to be powerful in other domains, such as image retrieval [23] and affective computing [10].

The remainder of this paper is organized as follows; in the next section, previous works related to automatic ontology generation are described. Section 3 explains the automatic tagging system and algorithms used in this work. Section 4 presents the experiments and results. Finally, in the last section, the paper concludes on the importance of the current research problem, and presents the next steps in our research.

2 Related Work

Although there are many ways of collecting experimental data for music information retrieval (MIR) research, the main challenges are the sparsity of the data, and the bias introduced by erroneous annotations. Besides, the cognitive processes underlying the representation and categorization of music are not yet fully understood, and it is often difficult to know what makes a tag accurate and what kinds of inaccuracies are tolerable [12, 9].

[2] www.last.fm

Last.fm is a popular online streaming service and social network which provides metadata assigned to songs or artists by users through an application programming interface (API). Social network users usually prefer to use the most frequent tags rather than by entering new tags in the system. Therefore, the obtained metadata may suffer from a popularity bias.

The most used classification systems for audio tagging are standard binary classifiers such as support vector machines (SVMs) and AdaBoost [26]. As supervised techniques, these classifiers rely on a training and a testing stage. Thereby, the classifier is engaged in predicting the musical tags of a testing dataset. Gaussian mixture model (GMM) is another well known technique that has been widely used in music tag prediction. The approach has shown to provide good semantic annotations for an acoustically diverse set of songs and retrieved relevant songs given a text-based query in [27]. In many studies, a time series of mel-frequency cepstral coefficient (MFCC) vectors are used as a music feature representation. MFCCs are a general purpose measure of the smoothed spectrum of an audio signal which primarily represent the timbral aspects of the sound. Although MFCCs are based on a simple auditory model and are common in the music and speech recognition world [5, 2]. The multi-layer perceptron (MLP) is one of the most commonly used neural networks. It can be used for classification problems, model construction, series forecasting and discrete control. For the forecasting problems, a backpropagation (BP) algorithm is normally used to train the MLP Neural Network [20, 19]. Since the MPL is very common in many research fields, and that neuro-fuzzy neural networks are based on the same learning framework, we have used this algorithm in our experiments, for comparison.

Parallel to this, there are ontologies in use today focusing on cases such as the classification of musical instruments [15]. For such sets of data, the primary organizational structure often involves spatial relationships; for example, object A connects to object B, object B is part of object A, object C is externally connected object B, object C is part of object A. One formalization of spatial relationships for the purpose of qualitative reasoning in ontological models is provided by Coalter and Leopold, in [4]. Fuzzy spatial reasoning is based on spatial relationships that provides a framework for modeling spatial relations in the fuzzy-set theory [24, 17, 6].

3 Audio Tagging System

The general architecture of the proposed audio tagging system is shown in Figure 1 and presented in the sections below.

3.1 Data Acquisition

For the data acquisition, two large databases were used: *i*) the Isophone database[3], [25] and *ii*) the Last.fm database. The Isophone database is based on the SoundBite plugin [16], which is available as iTunes and Songbird[4] plugins. The Sound-

[3] http://www.isophonics.net/

[4] `http://getsongbird.com/`

Bite plugin extracts features (MFCCs) from the entire user audio collection and stores them for further similarity calculations. The extracted features are also uploaded to a central server and expand dynamically the Isophone database.

The Isophone database uses MusicBrainz[5] identifiers as a source for unique identifiers. MusicBrainz is a comprehensive public community music metadata service. It can be used to identify songs or CDs, and provides valuable data about tracks, albums, artists and other related information. In order to associate the Isophone database to the MusicBrainz dataset, the GNAT[6] application is used, which implements a variant of the automated inter linking algorithm. In the metadata (tags) filtering process, MusicBrainz IDs of the tracks included in the Isophone database are matched against those of the Last.fm database by using Last.fm's AP. The collected tags were sorted out by their frequency of appearance within the Isophone database.

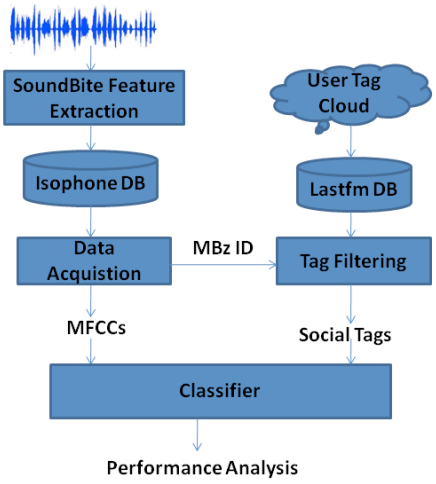


Fig. 1. Audio Tagging System

3.2 Classifiers

The classification is performed by using multi-layer perceptron and neuro-fuzzy systems which are supervised methods. Our goal is to associate an audio signal with various labels from a priori defined tag sets.

Multi-Layer Perceptron Neural Networks have been used in many different areas to solve pattern recognition problems. The multi-layer perceptron

[5] http://musicbrainz.org/
[6] http://www.sourceforge.net/projects/motools/

(MLP)[21] is one of the most common Neural Networks in use. It consists of two main computational stages: a feed-forward network and a backpropagation network. In the forward pass, input vectors are applied to the input nodes of the network, and at each node (neuron), the weighted sum of the input is computed. In the final stage of the forward pass, the set of outputs is produced as the actual output of the network. During the backward pass, the actual output of the network is subtracted from a desired output to produce an error signal, and the network weights are adjusted to move to the desired response according to the errors that are propagated backwards through the network. Fig. 2 shows the architecture of the Multi-Layer Perceptron used for deriving music tagging outputs from MFCCs.

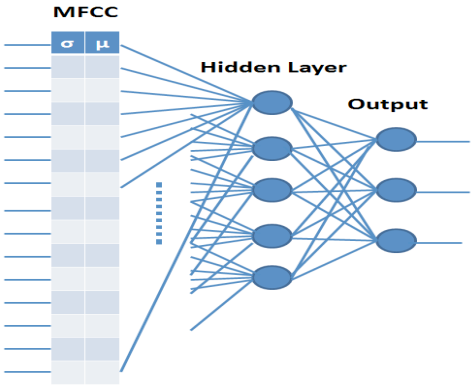


Fig. 2. Multi-Layer Perceptron for Music Tagging. σ and μ represent the variance and mean of the MFCCs time series, respectively

Neuro-Fuzzy Neuro-fuzzy (NF) systems [11] are a combination of neural networks and fuzzy logic [14] that merge the learning ability of neural networks and the reasoning ability of fuzzy logic. Automatic linguistic rule extraction is a typical application of neuro-fuzzy when there is little or no prior knowledge about the process. Figure 3 shows the architecture of a Neuro-Fuzzy network with two inputs and one output.

Considering the fuzzy sets of MFCC coefficients, the following linguistic rule set illustrates a simple fuzzy reasoning process. The MFCC coefficients are defined as the input variables, denoted $x_{1,1}, x_{1,2}, ...x_{i,j}$, where i and j refer to the rules and fuzzy sets, respectively. The rules can be expresses as follows:

$$\text{Rule } 1 : \overbrace{If\ x_{1,1} \text{ is } M_{1,1} \text{ and } x_{1,j} \text{ is } M_{1,j}}^{antecedent} \overbrace{\text{ then } y_1 \text{ is } y_d}^{consequent}$$
$$\vdots$$
$$\text{Rule } i : If\ x_{1,1} \text{ is } M_{i,1} \text{ and } x_{i,j} \text{ is } M_{i,j} \quad \text{then } y_i \text{ is } y_d$$

where M represents the fuzzy sets for the MFCC coefficients and y_d is the desired output provided based on music tags. In the fuzzification process, we used triangular symmetric membership functions. By acting on the parameters of the triangular membership functions, denoted a_{ij} and b_{ij}, it is possible to generate different types of functions (e.g. low, medium, high). Corresponding parameters of the membership function is defined below in Eq.1. Once the rules are determined, the inputs are fuzzified to obtain a membership degree, $\mu_{i,j}$, for each membership function of fuzzy sets, as follows:

$$\mu_{i,j} = \begin{cases} 1 - \dfrac{2 \mid x_j - a_{i,j} \mid}{b_j}, & a_{i,j} - \dfrac{b_{i,j}}{2} < x_j < a_{i,j} + \dfrac{b_{i,j}}{2} \\ 0 & \text{, otherwise} \end{cases} \tag{1}$$

Next, each satisfied fuzzy set's membership degree is used as an input to the fuzzy reasoning process which performs T-norm product operation. The consequent of a fuzzy rule assigns the entire rule to the output fuzzy set which is represented by a membership function that is chosen to indicate the related music tag. In the next layer the firing strengths of each rule are normalised. The normalised consequent fuzzy sets encompass many outputs, so it must be resolved into a single output value by a defuzzification method. In the defuzzification stage, the fuzzy sets which represent the outputs of each rule are combined into a single fuzzy set and distill a single output value from the set. The centre of gravity method which is one of the most popular defuzzification method has been used in the proposed approach to resolve the aggregated fuzzy set.

There are three types of parameters to be adapted in the learning stage which determine the parameter vector z:

$$z = (a_{11}, ..., a_{ij}, b_{11}, ..., b_{ij}, w_1, ..., w_i) \tag{2}$$

where a_{ij}, b_{ij} are the MFCC membership functions and w_i is the weight parameter that is used to tune the membership functions. The learning stage of the neuro-fuzzy approach uses neural nets learning system by optimising a criterion function (V) given by:

$$\nabla_z V = \left[\frac{\partial V}{\partial z_1}, ..., \frac{\partial V}{\partial z_i} \right] \tag{3}$$

where $-\nabla_z V$ is the gradient of V with respect to z. In order to tune the fuzzy set parameters, the weights and membership function's parameters need to be adjusted so as to minimize the error. Eq. (4) shows how to apply the

method of stochastic approximation on the criterion loss function to identify the parameters of the system. It is an iterative procedure given by:

$$z(t+1) = z(t) - \eta \nabla_z V[z(t)] \tag{4}$$

where z is the vector parameters to adapt and η is the predefined learning rate constant which specifies the computation speed of the learning task.

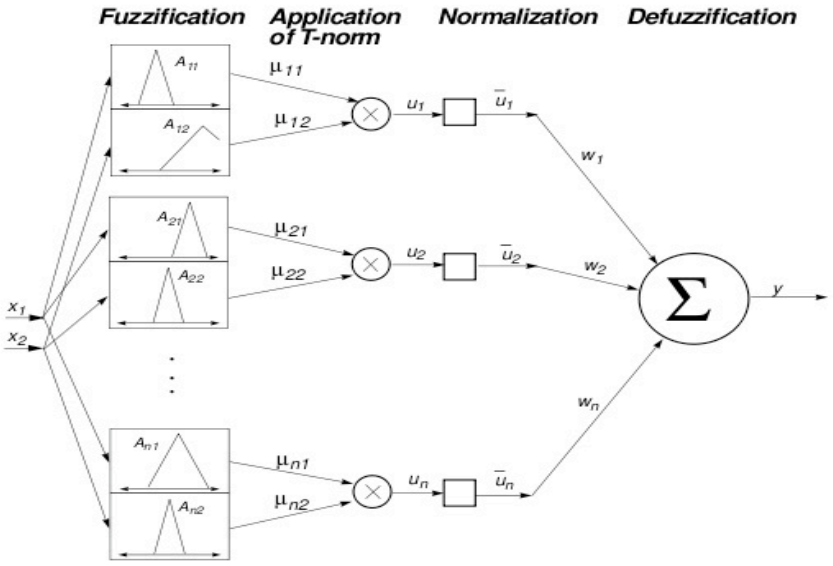


Fig. 3. Neuro-fuzzy system architecture (based on [7])

4 Experiments

Both of the neuro-fuzzy (NF) system and the multi-layer perceptron (MLP) neural network are based on the same network topologies and they were designed with multi-network system.

4.1 Dataset

The experimental dataset is a merge of Last.fm social tags for the Isophone database. In the experiments, 41 962 songs have been used out of 152 410 songs of the Isophone database. For each track we collected tags related to the five following categories: genre, mood, instrumentation, locale, and time-period. By summing up the subclasses associated with these tag categories, 32 tag subclasses

were considered in total (e.g. pop, chillout, guitar, american, 90s). For each given tag, 50% of the associated tracks were used for training, and 50% were used for testing. The repartition of tracks according to the various types of tags is given in Table 1. For each track, an audio feature vector of 40 values representing the mean and variances of 20 MFCCs is computed, as in [25].

Genre	Data %	Instrumentation	Data %	Mood	Data %	Locale	%	Time-Period	Data %
Pop	38.52	Electronic	11.51	Dance	7.75	American	20.69	00s	14.67
Alter. Rock	26.45	Acoustic	11.48	Relax	6.14	French	1.92	90s	20.91
Classic Rock	25.70	Guitar	9.20	Fun	4.81	Swedish	1.10	80s	15.22
Electronica	12.18	Piano	10.66	Melancholic	17.40			70s	14.55
Punk	13.92	Vocal	10.14	Party	13.46			60s	10.20
Hard Rock	13.70			Romantic	14.32				
Jazz	13.74			Atmospheric	7.77				
Blues	12.70								
Ambient	9.41								
Trip Hop	5.35								
Soul	10.30								
Metal	11.00								
Total	**88.13**		**36.87**		**51.13**		**23.65**		**57.89**

Table 1. Repartition of tracks in the experimental data set according to genre, instrumentation, mood, locale, and time-period

4.2 Analysis parameters

The number of iterations in the neuro-fuzzy and MLP algorithms were identified according to the lowest point on the mean square error curves obtained in the training stage. The best learning rate ($\eta = 0.6$) was determined empirically. For each tag, the structure of the MLP consisted of 40 input nodes, 20 hidden nodes, and 1 output node. In calculating the hidden and output units of the MLP the *tanh* function was used as the activation function. In the neuro-fuzzy system each network was created with the 40 inputs and 1 output rule set. Three membership functions have been used for each fuzzy set (low, medium, and high). Both algorithms comprised 32 different networks in total.

4.3 Results

In order to evaluate the performance of these algorithms, standard evaluation metrics (precision [P], recall [R], F-measure [F]) have been used [18].

The results are shown in Table 2. On overall, the neuro-fuzzy system achieved an F-measure of 46% in the identification of a large number of music tags (32). The multi-layer perceptron's overall F-measure was 21% that is lower by 25% points in comparison with that of the NF method. The better results obtained for the labels "vocal", "melancholic", "metal", "classic rock", and "60s". The labels "party", "atmospheric", "romantic", "fun" obtained the lowest performance in this experiment. This is probably due to the fact that other factors than timbre (as modeled by the MFCCs) are involved to characterise these genres and emotion-eliciting situations (e.g. rhythm for party music is deemed to be very important). The results indicated that neuro-fuzzy systems performed much better than the multi-layer perceptron on large-scale experiments.

		P		R		F	
		NF	MLP	NF	MLP	NF	MLP
Genre	Pop	**0.66**	0.57	**0.52**	0.46	**0.58**	0.51
	Alter. Rock	**0.65**	0.55	**0.51**	0.32	**0.57**	0.41
	Classic Rock	**0.70**	0.58	**0.54**	0.32	**0.61**	0.41
	Electronica	**0.64**	0.57	**0.41**	0.22	**0.50**	0.31
	Punk	**0.62**	**0.62**	**0.35**	0.29	**0.45**	0.39
	Hard Rock	**0.68**	0.54	**0.48**	0.20	**0.56**	0.29
	Jazz	0.67	**0.73**	**0.41**	0.34	**0.51**	0.46
	Blues	**0.62**	0.45	**0.34**	0.08	**0.44**	0.14
	Ambient	**0.62**	0.49	**0.29**	0.19	**0.40**	0.27
	Trip Hop	**0.67**	0.40	**0.36**	0.04	**0.47**	0.06
	Soul	**0.64**	0.45	**0.36**	0.13	**0.46**	0.21
	Metal	**0.73**	0.61	**0.57**	0.31	**0.64**	0.41
	Average	**0.65**	**0.54**	**0.42**	**0.24**	**0.51**	0.32
Instrumentation	Electronic	**0.64**	0.36	**0.44**	0.07	**0.52**	0.11
	Acoustic	**0.53**	0.46	**0.23**	0.10	**0.32**	0.17
	Guitar	**0.54**	0.32	**0.24**	0.06	**0.33**	0.11
	Piano	**0.56**	0.55	**0.20**	0.02	**0.29**	0.04
	Vocal	**1.00**	0.43	**1.00**	0.04	**1.00**	0.07
	Average	**0.65**	0.42	**0.42**	0.05	**0.49**	0.10
Mood	Dance	**0.53**	0.31	**0.20**	0.04	**0.30**	0.07
	Relax	**0.51**	0.39	**0.14**	0.03	**0.22**	0.05
	Fun	0.31	**0.36**	**0.07**	0.01	**0.12**	0.02
	Melancholic	**1.00**	0.64	**1.00**	0.32	**1.00**	0.42
	Party	0.21	**0.53**	0.02	**0.18**	0.04	**0.27**
	Romantic	0.34	**0.44**	**0.05**	0.03	**0.08**	0.06
	Atmospheric	0.37	**0.45**	0.07	**0.11**	0.11	**0.17**
	Average	**0.46**	0.44	**0.22**	0.10	**0.26**	0.15
Locale	American	**0.58**	0.42	**0.36**	0.06	**0.44**	0.10
	French	**0.67**	0.15	**0.40**	0.04	**0.50**	0.06
	Swedish	**0.64**	0.26	**0.47**	0.09	**0.54**	0.13
	Average	**0.63**	0.27	**0.41**	0.06	**0.49**	0.09
Time-Period	00s	**0.56**	0.45	**0.30**	0.11	**0.39**	0.18
	90s	**0.63**	0.44	**0.45**	0.11	**0.52**	0.17
	80s	**0.65**	0.52	**0.43**	0.14	**0.52**	0.23
	70s	**0.63**	0.50	**0.45**	0.10	**0.53**	0.17
	60s	**0.72**	0.56	**0.56**	0.12	**0.63**	0.20
	Average	**0.63**	0.49	**0.43**	0.11	**0.51**	0.19
Overall		**0.61**	0.47	**0.38**	0.14	**0.45**	0.20

Table 2. Performance of the neuro-fuzzy (NF) system and multi-layer perceptron (MPL) network in the classification of five music tag classes: genre, instrumentation, mood, locale, and time-period

5 Discussion

Reasonably good performance were obtained for the neuro-fuzzy system in the case of genre, time period, and location, considering the large number of classes (32) in these experiments. However the results were poor for the mood and instrumentation labels showing the need to refine the features and/or classification framework. Research on music emotion recognition has shown that the regression approach applied to arousal/valence values outperformed the classification approach applied to categorical labels [13]. Research on polyphonic musical instrument recognition is still in its early days [8], and it is not surprising to obtain low recognition accuracy for the instrumentation since the MFCCs only capture the timbre of the music at a "macro" level (globally). It should also be noted that label inaccuracies in the social data may have affected the results for both classifiers. However as previously mentioned the main goal of the study was to compare the relative performance of the NF and MLP methods with regards to the promising application of NF systems in automatic ontology generation.

Our study provides a framework for future studies to assess systems using the Isophone dataset. Although no means are offered for automatically extracting and proposing axioms to ontology engineering in this study, future work will investigate the identifications of the relationships between different conceptual entities as in [4]. As an example of the future use of ontologies on music annotation systems, it is also worth to mention a recent study proposed by Wang et al.[28] in which an ontology-based semantic reasoning is used to bridge content-based information with web-based resources. The authors pointed out that the proposed ontology-based system outperformed content-based methods and significantly enhanced the mood prediction accuracy.

6 Conclusion

Our research is motivated by the fact that, current ontology designs have inflexible structure and have not been used with any automated learning system which leads to a danger to fossilise the current knowledge representation by static ontologies. Preliminary analyses were conducted with a neuro-fuzzy (NF) system and a multi-layer perceptron (MLP) neural network in a music-tag annotation task. The results showed that NF outperformed MLP by 25% points in F-measure, which indicated that fuzzy systems are promising classifiers for audio content-based ontology construction. In our future work, our study will continue towards the automatic ontology generation by using fuzzy spatial reasoning systems.

References

1. G. Antoniou and F. can Harmelen. *Semantic Web Premier 2nd Edition.* Massachusetts Institute of Technology, 2008.

2. J.-J. Aucouturier and F. Pachet. Music similarity measures: Whats the use? *Proceedings of the 3rd International Symposium on Music Information Retrieval*, pages 157–163, 2002.
3. D. S. C. Catton. The use of named graphs to enable ontology evolution. *W3C Workshop on the Semantic Web for Life Sciences*, 2004.
4. A. Coalter and J. L. Leopold. Automated ontology generation using spatial reasoning. *KSEM'10 Proceedings of the 4th international conference on Knowledge science, engineering and management.*, pages 482–493, 2010.
5. D. P. W. Ellis. Classifying music audio with timbral and chroma features. *The International Society of Music Information Retrieval*, 2007.
6. A. Esterline, G. Dozier, and A. Homaifar. Fuzzy spatial reasoning. *Fuzzy spatial reasoning. In:Proc. of the 1997 Int. Fuzzy Systems Association Conference*, 1997.
7. J. Godjevac. Comparative study of fuzzy control, neural network control and neuro-fuzzy control. *Technical Report*, 1995.
8. P. Herrera-Boyer, A. Klapuri, and M. Davy. Automatic classification of pitched musical instrument sounds. In A. Klapuri and M. Davy, editors, *Signal Processing Methods for Music Transcription*, pages 163–200. Springer, 2006.
9. M. D. Hoffman, D. M. Blei, and P. R. Cook. Easy as cba: A simple probabilistic model for tagging music. In *Proceedings of the 10th International Society for Music Information Retrieval Conference*. International Society for Music Information Retrieval, October 2009.
10. S. V. Ioannou, A. T. Raouzaiou, V. A. Tzouvaras, T. P. Mailis, K. C. Karpouzis, and S. D. Kollias. Emotion recognition through facial expression analysis based on a neurofuzzy network. *Neural Networks*, 18:423–435, 2005.
11. E. M. Jyh-Shing Roger Jang, Chuen-Tsai Sun. *Neuro-Fuzzy and Soft Computing*. Prentice-Hall, Inc, 1997.
12. J. H. Kim, B. Tomasik, and D. Turnbull. Using artist similarity to propagate semantic using artist similarity to propagate semantic information. In *10th International Society for Music Information Retrieval Conference*, 2009.
13. Y. E. Kim, E. M. Schmidt, R. Migneco, and B. G. Morton. Music emotion recognition: a state of the art review. *11th International Society for Music Information Retrieval (ISMIR) Conference*, pages 255–266, 2010.
14. G. J. Klir and B. Yuan. *Fuzzy Sets and Fuzzy Logic Theory and Applications*. Prentice Hall PTR, 1995.
15. S. Kolozali, G. Fazekas, M. Barthet, and M. Sandler. Knowledge representation issues in musical instrument ontology design. In *12th International Society for Music Information Retrieval Conference (ISMIR)*, Miami, Florida (USA), 2011.
16. M. Levy and M. Sandler. Lightweight measures for timbral similarity of musical audio. In *Proc. of the 1st ACM workshop on Audio and Music Computing multimedia*, New York, 2006.
17. C. Liu, D. Liu, and S. Wang. Fuzzy geopspatial ontology model and its application in geospatial semantic retrieval. *Journal of Convergence Information Technology*, 6, 2011.
18. C. D. Manning, P. Raghavan, and H. Schutze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.
19. S. Mirjalili and A. Sadiq. Magnetic optimization algorithm for training multi layer perceptron. *Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on*, pages 42 – 46, 2011.
20. R. Neumayer. Musical genre classification using a multi layer perceptron. *Proceedings of the 5th Workshop on Data Analysis (WDA'04)*, pages 51–66, 2004.

21. S. Rajashekaran and G. Vijayalksmi. *Neural Networks, Fuzzy Logic and Genetic Algorithms: Synthesis and Applications*. Springer, 2004.
22. V. Sebastien, D. Semastien, and N. CONRUYT. An ontology for musical performances analysis. In *2010 Fifth International Conference on Internet and Web Applications and Services*, pages 538–543, Barcelona, May 2010.
23. N. Singhai and S. K. Shandilya. A survey on: Content based image retrieval systems. *International Journal of Computer Applications*, 4(2):22–26, 2010.
24. U. Straccia. Towards spatial reasoning in fuzzy description logics. *The annual IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 512–517, 2009.
25. D. Tidhar, G. Fazekas, S. Kolozali, and M. Sandler. Publishing music similarity features on the semantic web. *Proceedings of the 10th International Society for Music Information Retrieval Conference*, 2009.
26. K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas. Multi-label classification of music into emotions. In *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR)*, 2008.
27. D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet. Semantic annotation and retrieval semantic annotation and retrieval of music and sound effects. In *IEEE Transactions on Audio, Speech, and Language Processing*, volume 16, February 2008.
28. J. Wang, X. Anguera, X. Chen, and D. Yang. Enriching music mood annotation by semantic association reasoning. *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1445–1450, 2010.
29. T. Weyde and K. Dalinghaus. A neuro-fuzzy system for sequence alignment on two levels. *Mathware & softcomputing*, 11:197–210, 2004.

Oral session 2:
3D Audio and Sound Synthesis

A 2D Variable-Order, Variable-Decoder, Ambisonics based Music Composition and Production Tool for an Octagonal Speaker Layout

Martin J. Morrell[1] and Joshua D. Reiss[1] *

Centre for Digital Music, Queen Mary University of London, London, UK
martin.morrell@eecs.qmul.ac.uk

Abstract. This paper introduces a music production/composition tool for the spatialisation of sound sources played over an octagonal loudspeaker layout. The tool is based on Ambisonics theory, but does not produce any intermediary B-Format signals. The novel aspects of the tool is that it allows for variable-order and variable-decoder attributes on a per sound source basis. This allows creative control over the sounds' localisation sharpness. Distance including inside the speaker layout source placement and reverberation attributes can be assigned to each sound source to create a final spatial mix. The theory of variable-order, variable-decoder Ambisonics is discussed and the implementation aspects presented. The authors aim to bridge the gap between theory and usage of Ambisonics.

Keywords: Ambsionics, variable-order, variable-decoder, octagon, spatial audio, 2D, 3D.

1 Introduction

Ambisonics is a spatialisation technique for recording, panning and reproducing two and three-dimensional sound sources. Work on Ambisonics is often very much theoretical research and at other times is artists using Ambisonics ready tools to produce work. However the latter usually relies on using software such as Max/MSP to create custom software to then create artistic work or the use of plugins for digital audio workstations that are not well equipped for handling the amount of channels that Ambisonics can produce or the eventual speaker feeds needed. In this paper the authors present the theory behind the creation of a new tool that allows users to send audio from current digital audio workstation projects to be spatialised around an eight speaker octagonal layout. The tool offers some new novel features discussed in-depth in the paper to create variabe-order and variable-decoder based Ambisonics-esque signals. The spatialisation tool is controlled via standard midi protocol and the parameters stored in the same digital audio workstation project.

* This research was supported by the Engineering and Physical Sciences Research Council [grant number EP/P503426/1].

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

2 Ambsionics Background

Michael Gerzon led the original Ambsionics development team in the 1970s and wrote papers on the subject throughout his life [8–10]. Further work has been done to expand Ambisonics into Higher Order Ambisonics [2–4, 12] and to develop decoders, speaker layouts and evaluation of systems [1, 6, 7, 11, 13, 14, 18]. The basis of Ambisonics is to represent a three-dimensional auditory scene as a field representation that can later be reconstructed for any user loudspeaker layout. An Ambisonics representation is based on a fixed order that is linked to the localisation attributes of sound sources. Ambisonics theory is based on spherical harmonics calculated from legendre polynomials.

$$Y_{mn}^{\sigma(N2D)}(\theta,\phi\,) = \sqrt{2}\hat{P}_{mn}(\sin\phi) = \begin{cases} \cos n\theta n & \geq 0 \\ \sin n\theta & n < 0 \end{cases} . \tag{1}$$

$$\hat{P}_{mn}(\sin\phi) = \sqrt{(2-\phi_{0,n})\frac{(m-n)!}{(m+n)!}}P_{mn}(\sin\phi) \; . \tag{2}$$

The above equations use the N2D normalisation scheme. Several schemes exist for Ambisonics and affect the maximum gain of each spherical harmonic. When these are applied to a monaural sound source a sound field representation is created and is known as B-Format. The 2D representation is based only on the angular value θ as $\phi = 0$. The spherical harmonic expansion of the sound field is truncated to a finite representation known as the Ambisonic order M and each prior order m is included, $0 \leq m \leq M$. For each included order m the degrees calculated are $n = \pm m$. Where the total amount of harmonics in the sound field representation is $2M + 1$.

Once encoded, Ambisonics material can be played back over various different loudspeaker layouts using a suitable decoder. The minimum number of loudspeakers to correctly reproduce 2D Ambisonics is $2M + 2$ [17]. For a regular layout, i.e. one that has the loudspeakers equally spaced, the angular separation is simply $360^o/L$ where L is the number of loudspeakers for 2D reproduction. For a regular layout the decoder matrix can be calculated by using the Moore-Penrose pseudo-inverse matrix of the spherical harmonics (equal to the source material and appropriate for the amount of loudspeakers) at each loudspeaker position.

$$pinv \begin{pmatrix} Y_{(0,0)}(spk1) & Y_{(1,-1)}(spk1) & Y_{(1,1)}(spk1) & \dots Y_{(M,m)}(spk1) \\ Y_{(0,0)}(spk2) & Y_{(1,-1)}(spk2) & Y_{(1,1)}(spk2) & \dots Y_{(M,m)}(spk2) \\ Y_{(0,0)}(spk3) & Y_{(1,-1)}(spk3) & Y_{(1,1)}(spk3) & \dots Y_{(M,m)}(spk3) \\ \vdots & \vdots & \vdots & \ddots \vdots \\ Y_{(0,0)}(spkN) & Y_{(1,-1)}(spkN) & Y_{(1,1)}(spkN) & \dots Y_{(M,m)}(spkN) \end{pmatrix} . \tag{3}$$

The given pseudo-inverse decoder results in the standard, rV, decoder matrix. Gerzon specified criteria for low and high frequencies reproduction known as rV

and rE vectors [8, 9, 11]. To create a decoder that maximises the rE vector the decoder is then multiplied with gains $g'm$ based on each component's order and the system order.

$$g'm = P_m(\text{largest root of } P_{M+1}) \ . \tag{4}$$

Furthermore the decoding can be changed to what is known as in-phase decoding so that there are no negative gains used to create the sound's directionality.

$$g'm = \frac{M!}{(M+m)!(M-m)!} \ . \tag{5}$$

Ambisonics can be seen as creating a polar pattern of M^{th} order in the direction of the sound source where the polar pattern is sampled by discrete loudspeaker positions. By increasing the amount of loudspeakers the resolution of the polar pattern is increased. In turn, by increasing the order, the directionality is increased and by using different decoders as described above, the rear-lobe is altered.

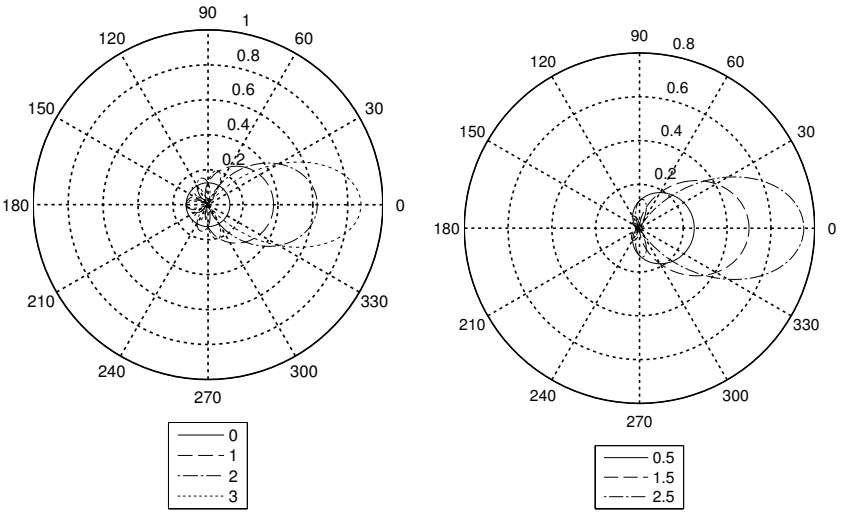
3 Variable-Order and Variable-Decoder Concept

In this section the authors present the novel idea of variable-order and variable-decoder Ambisonics. This concept allows for varying the reproduced polar pattern, and therefore the sharpness of localisation, by setting the order used to a non-integer value. Further to this, the idea of a variable-decoder is discussed that can alter the amount of rear lobe of the sampled polar pattern. The two variables are linked but not interchangeable. The order alters the width of the main lobe, whilst altering the amount of and gain of, the rear lobes. The decoder alters the gain of rear lobes whilst consequently altering the width and gain of the main lobe.

3.1 Variable-Order

The result of encoding a monaural sound source to Ambiosnics B-Format and then decoding it for a loudspeaker layout is equivalent to applying a gain to the monaural sound and sending it to each loudspeaker. Therefore in the authors' approach the audio signal is not converted to B-Format. Instead the gains are calculated numerically and applied based on the octagon layout.

The variable-order is created by calculating the decoders, of same type, for each order. Since we are dealing with an octagonal layout the orders used are 0 through 3. The spherical harmonic values are calculated for all orders for the sound source location θ and speaker gains obtained. By using linear algebra the variable-order can be created by a mixture of 0^{th} and 1^{st}, 1^{st} and 2^{nd}, and 2^{nd} and 3^{rd} speaker gains. Figure 1 a) shows the sampled polar pattern for the whole orders. Figure 1 b) shows the half orders using the variable-order approach. As



(a) Orders 0, 1, 2 and 3 (b) Orders 0.5, 1.5 and 2.5

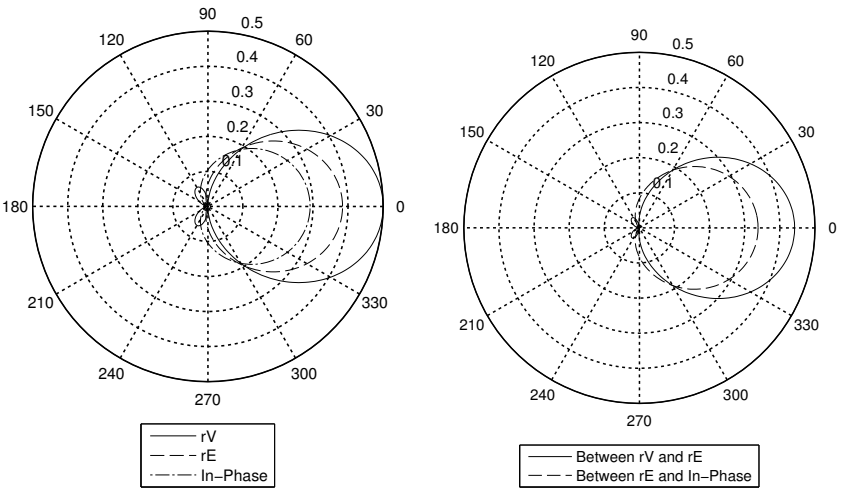
Fig. 1: The reproduced polar pattern of a sound source at $\theta = 0^o$ for Ambisonic orders 0 through 3 are shown in a). The half orders of 0.5, 1.5 and 2.5 are shown in b).

can be expected the polar pattern of half orders is directly between the whole orders. The variable-order approach can be used to create the polar pattern of any decimal value order representation. For an Ambisonics representation the gain of all loudspeakers must equal 1. This has been calculated to be true, but from simple algebra if the two speaker feeds both equal one and the weighting applied equal one, then so must the resultant equal one. This fact is important so that a sound source does not experience an overall gain boost when the variable-order is used as a creative feature.

3.2 Variable-Decoder

Three types of Ambisonics decoders have been presented in section 2 and each is used for a specific purpose. However these decoders offer an aspect of creativity over being able to manipulate the rear lobe of the polar pattern, thus altering the shape of the sound sources' polar pattern.

As for the variable-order concept, the variable-decoder can be calculated in the same manner. By using a weighted ratio that equals 1 of two types of decoder, a variable pattern can be created. The weighting is done between rV and rE decoders and the rE and in-phase decoders. This is because the rE polar pattern lies between the basic and in-phase patterns.



(a) rV, rE and In-Phase decoders (b) Variable-decoders

Fig. 2: The three standard decoder types for order 1.5 are shown on the left and the intermediate decoders on the right.

Figure 2 shows the three decoders for order 1.5 on the left and the decoders half way between the rV snd rE decoders and the rE and in-phase decoders. The variable-decoder lies at the given ratio between the standard decoders.

3.3 Observations

The proposed methodology creates a set of variable-order, varidable-decoder loudspeakers signals for an octagon arrangement of loudspeakers. The end result is sampling at regular intervals of a third oder polar pattern [5]. The resultant gain g_L for loudspeaker at position θ_L can be calculated by equation 6. The sum of the gain of each order must equal one.

$$g_L = a_0 + a_1 \cos(\theta + \theta_L) + a_2 \cos(2(\theta + \theta_L)) + a_3 \cos(3(\theta + \theta_L)) \; . \qquad (6)$$

Therefore the variable-order is equivalent to increasing the next order gain whilst the ratio of the prior orders' gains remains the same. The variable-decoder, is like altering the ratio between the a_0 and a_1 gain coefficients thus changing the base polar pattern, as well as altering the ratio between higher orders.

4 Composition/Production Tool Implementation

The tool to use the variable-order, variable-decoder methodology has been implemented in the Max/MSP 5 software environment for Mac OSX. The tool is

designed to receive audio signals from digital audio workstations, e.g. via Jack or Soundflower, for a total of 16 monaural and 4 stereophonic signals. The controls for each channel are sent via midi commands which can be stored in a digital audio workstation project. The authors built User Control Panels for this function for the Cubase/Nuendo environment, but VSTs, AUs or other midi capable software can be used to control the settings for each sound source. The premise for this is that no extra saved data is needed that cannot be stored in a common audio project.

Figure 3 shows the user interface for the tool. The only user definable parameters on the interface are On/Off, midi driver, audio driver and where to save a recorded file. The interface has eight LED style meters for monitoring the signal level going to each loudspeaker so that distortion can be avoided. Since users may not always have an eight speaker layout available, a binaural (over headphones) mix is simultaneously available.

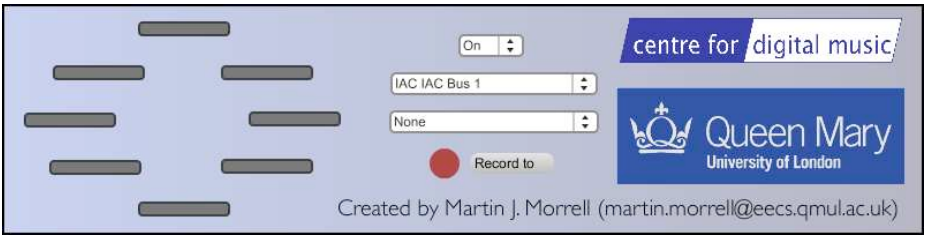


Fig. 3: The user interface for the variable-order, variable-decoder spatialisation tool.

4.1 Tool Features

The novel features in this tool have already been presented in this paper, however, there are some other features that are note worthy. This includes the handling of distance and reverberation.

Distance Distance is a user definable parameter and is accomplished by gain manipulation only. No delay has been included since for music purposes pitch shifting of sound sources will affect the overall tonal effect of the work, alter the speed and therefore ensemble timing of the music and finally can include zipper noise. The $1/r$ inverse law is used to implement the gain change at sources greater than 1.0 where the maximum value is 10. Since the roll off of $1/r$ simulates anechoic conditions, the feature is given as for creative not real-world application. For sources that are placed inside the speaker layout the distance calculation changes to $1 + \cos(90^o r)$ so that infinite gain is not reached. The maximum gain at the central position is 2.0, or approximately +6dB.

Inside Panning Sound sources that have a distance less than 1.0 are placed inside the loudspeaker array. This is done by altering the polar patterns. If the order of reproduction is 1 then this is the same as cancelling out the 1st order spherical harmonics and doubling the zeroth order spherical harmonics [15]. For the case of third order 2D Ambisonics, the maximum allowed in this tool, the inside panning function is expanded. The result is that even orders are doubled and odd orders are cancelled out. This again is all done as numerical and not audio calculations. Figure 4 shows the polar pattern change going from 1.0 to 0.0. The result is strong lobes from opposite poles giving the psychoacoustic illusion of being in the centre of the array.

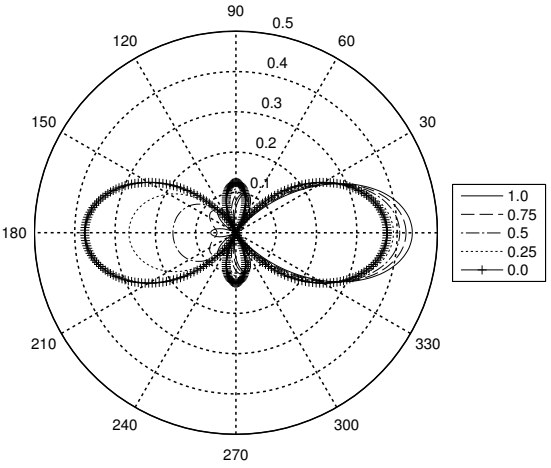


Fig. 4: The change in polar pattern exerted by a third order sound source as it is moved from a distance of 1.0 to 0.0 to be placed in the middle of the speaker array.

Reverberation Reverberation is produced in the tool by transforming the sound source into B-Format and processing it through either the Wigware VST reverberation plugin [18] based on the freeverb algorithm or using a convolution plugin using B-Format impulse responses, such as those freely available [16].

5 Conclusion

The authors have presented a novel approach to implementing the theory of Ambisonics that does not use the intermediary B-Format representations for a fixed octagonal loudspeaker layout. These conditions however mean that composers/artists do not need to worry about designing speaker layouts. Furthermore by fixing the speaker layout of the tool, calculations are done numerically,

and a variable-order and decoder is created for each sound source. The result is being able to mix orders to create a composer defined rather than technologically defined sound field that the listener hears. Work is being undertaken to use this tool for an original composition.

References

1. Benjamin, E.: Ambisonic loudspeaker arrays. In: Audio Engineering Society Convention 125 (10 2008)
2. Daniel, J.: Reprèsentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimèdia. Ph.D. thesis, l'Universite Paris 6 (2000)
3. Daniel, J.: Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new ambisonic format. In: Audio Engineering Society Conference: 23rd International Conference: Signal Processing in Audio Recording and Reproduction (5 2003)
4. Daniel, J., Moreau, S.: Further study of sound field coding with higher order ambisonics. In: Audio Engineering Society Convention 116 (5 2004)
5. Eargle, J.: The Microphone Book. Focal Press (2001)
6. Furness, R.K.: Ambisonics-an overview. In: Audio Engineering Society Conference: 8th International Conference: The Sound of Audio (5 1990)
7. Furse, R.W.: Building an openal implementation using ambisonics. In: Audio Engineering Society Conference: 35th International Conference: Audio for Games (2 2009)
8. Gerzon, M.A.: Periphony: With-height sound reproduction. J. Audio Eng. Soc 21(1), 2–10 (1973)
9. Gerzon, M.A.: Practical periphony: The reproduction of full-sphere sound. In: Audio Engineering Society Convention 65 (2 1980)
10. Gerzon, M.A., Barton, G.J.: Ambisonic decoders for hdtv. In: Audio Engineering Society Convention 92 (3 1992)
11. Heller, A., Lee, R., Benjamin, E.: Is my decoder ambisonic? In: Audio Engineering Society Convention 125 (10 2008)
12. Käsbach, J., Favrot, S.: Evaluation of a mixed-order planar and periphonic ambisonics playback implementation. In: Forum Acusticum 2011 (2011)
13. Menzies, D.: W-panning and o-format, tools for object spatialization. In: 22nd International Conference on Virtual, Synthetic and Entertainment Audio (June 2002)
14. Moore, D., Wakefield, J.: The design of ambisonic decoders for the itu 5.1 layout with even performance characteristics. In: Audio Engineering Society Convention 124 (May 2008)
15. Morrell, M.J., Reiss, J.D.: A comparative approach to sound localization within a 3-d sound field. In: Audio Engineering Society Convention 126 (5 2009)
16. Murphy, D.T., Shelley, S.: Openair: An interactive auralization web resource and database. In: Audio Engineering Society Convention 129 (November 2010)
17. Poletti, M.A.: Three-dimensional surround sound systems based on spherical harmonics. J. Audio Eng. Soc 53(11), 1004–1025 (2005), `http://www.aes.org/e-lib/browse.cfm?elib=13396`
18. Wiggins, B.: Has ambisonics come of age? In: Proceedings of the Institute of Acoustics. vol. 30. Pt. 6 (2008)

Perceptual characteristic and compression research in 3D audio technology

Ruimin Hu, Shi Dong, Heng Wang, Maosheng Zhang, Song Wang, Dengshi Li

National Engineering Research Center for Multimedia Software

School of Computer Science, Wuhan University, Wuhan, 430072, China

hrm1964@163.com, edisonds@gmail.com, wh825554@163.com, eterou@163.com,

wangsongf117@163.com, reallds@126.com

Abstract. The 3D audio coding forms a competitive research area due to the standardization of both international standards (i.e. MPEG) and localized standards (i.e. Audio and Video Coding Standard workgroup of China, AVS). Perception of 3D audio is a key issue for standardization and remains a challenging problem. Besides current solutions adopted from traditional audio engineering, we are working for an original 3D audio solution for compression. This paper represents our initial results about 3D audio perception include directional measurement of Just Noticeable Difference (JND) and Perceptual Entropy (PE). We also represent the possible applications of these results in our future researches.

Keywords: 3D audio, perceptual audio processing, audio compression

1 Introduction

With the current trend of 3D movies and the popularization of 3DTV, 3D audio and video technology has become a research topic in multimedia technology. To provide the audience with a more immersive and integrated audio-visual experience, audio must work collaboratively with 3D video to provide three dimensional sound effects. However, existing 3DTV and 3D movie systems usually adopt conventional stereo audio and surround sound technology, which only provides very limited sound localization ability and envelopment in horizontal plane. Although there is not a generally acknowledged definition for 3D audio, it is widely accepted that 3D audio must have the following characteristics; localization of sound image in arbitrary direction in 3D space, realizing the distance perception of sound and giving a improved feeling of audio scene. Nowadays two types of technology are able to satisfy the requirement of 3D audio, one is based on physical principles and aims at reconstructing the original sound field, the other is based on principle of human perception and aims at giving the listener a virtual sound image. Wave Field Synthesis (WFS), Ambisonics and 22.2 multichannel systems are three typical 3D audio systems following those principles.

This paper is arranged as follows. In section 2 an introduction to the three 3D audio systems is presented and the existing problems are discussed, where we conclude the complexity of the 3D systems and efficiency of the signal compression will be two problems for the popularization of 3D audio. In section 3 we present our related work in 3D audio technology, including hearing mechanism and signal

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

compression research. More specifically, we investigate the JND of the direction perception cues for human in horizon plane. This is useful in simplification the 3D audio recording and playback systems, and removing the redundant perceptual information in 3D audio signals. In section 4 the development trends of 3D audio and our future work are discussed.

2 Brief view of typical 3D audio systems

2.1 Wave Field Synthesis (WFS)

a. The Principle of Wave Field Synthesis

The concept of WFS was introduced by Berkhout in 1988[1], its physical theory can date back to Huygens principle which suggests that a wave which propagates from a given wave front can be considered as emitted either by the original sound source or by a secondary source distribution along the wave front [2]. To reconstruct the primary sound field, the distribution of secondary source can replace primary source. The concept was later developed by Kirchhoff and Rayleigh, and the Kirchhoff-Helmholtz integral they proposed can be interpreted as follows: if appropriately secondary sources are driven by the values of the sound pressure and the directional pressure gradient caused by the virtual source on the boundary of a closed area, then the wave field within the region is equivalent to the original wave field[3]. By adding a degree of freedom to the secondary source distribution, Kirchhoff-Helmholtz generalized Huygens principle.

b. Realization of WFS

According to the above theory, WFS reproduces the primary sound field in time and space by making using of small and individually driven loudspeakers array, and can recover the spatial image precisely in the half space of receiving end from loudspeaker arrays[4].

But there is some limit for WFS in application. WFS needs a continuous, closed surface and a large number of idealized loudspeakers, but in practice there is only a discontinuous loudspeaker array. According to spatial nyquist sampling Theorem, if the interval between loudspeakers is less than half the wave length of a sound wave, aliasing will not occur[5].

So according to spatial nyquist sampling Theorem, WFS can be realized by limited and discrete loudspeakers within a certain frequency range. For example, limited line loudspeaker with even intervals can reconstruct sound field in 2D horizontal plane[6]. In the recording stage, the listening area is surrounded by a microphone array. The microphone array consists of pressure and velocity microphones, which record the primary sound field of external sound sources. In the reconstruction stage, the microphones will be replaced by the loudspeakers. Each loudspeaker is driven by signal recorded by the corresponding microphone. The geometric shape of the microphone array and loudspeaker are the same[7].

2.2 Ambisonics

a. The principle of Ambisonics

Ambisonics emerged in the 1970's and the main contributor is Gerzon [8] The principles of Ambisonics are as follows. A certain wave (sound field) can be expand on a sphere in sphere coordinate system by spherical harmonic functions. At the opposite end, superposition of spherical harmonic functions can rebuild a wave (sound field). There are $n=2m+1$ spherical harmonic functions at every order m of Ambisonics, a 3D system of M order consists of all spherical harmonic functions at every order m ($0 \leqslant m \leqslant M$), total channel number N satisfies $N=(M+1)^2$.

b. Two simple format of Ambisonics

The first format of Ambisonics proposed by Gerzon is B format, which displays an omnidirectional sound field by four channels: W, X, Y, Z [9]. Traditional monophony and stereophony can be seen as the subsystems of Ambisonics[10]. Sound location in horizontal plane is realized using three channels W, X, Y, and the fourth channel Z is used for reconstructing height information. Channel W is a pressure signal, and X, Y, Z are directional signal. B-format is used in studio and professional application.

The second format of Ambisonics is UHJ system which can convert directional sound into two or more channels and solve the incompatibility problem of four channels Ambisonics with monophony, stereophony [11][12]. The coding scheme provided by UHJ can be used in broadcasting, digital audio recording[13].

c. Playback technology of Ambisonics

According to the principle of Ambisonics, the decomposition of a sound field requires the expansion of infinite order spherical harmonic functions. But in practical application, limited order truncation of spherical harmonic functions expansion is necessary. B-format is one order expansion. Ambisonics was expanded to high order in the 1990's, the sweet point was enlarged to an area. High order Ambisonics promotes sound location with the price of more channels and loudspeakers. We can get better reconstruction quality using higher order Ambisonics. The encoding process of Ambisonics is to preserve the result of spherical harmonic functions multiplying the signal picked up by microphones. The decoding process is to calculate a group of loudspeaker signals according to the rebuilt sound field that must be equal to the primary sound field at listening point. This can be done by solving the inverse matrix which consists of spherical harmonic functions that are associated with locations of loudspeakers.

2.3 22.2 multichannel sound systems

a. Fundamentals of multichannel sound systems

The research of spatial hearing and sound source localization indicates that there are slight time and level differences between two ears when spatial sound signals arrive at the ears. For the estimation of direction and distance of sound source, the difference between the two ears signals is most relevant. Actually these differences, called binaural cues, are Interaural Time Difference (ITD) and Interaural Level

Difference (ILD). ILD and ITD indicate the level difference and time difference between left and right ears respectively [14].

b. Stereo, 5.1 surround sound and 22.2 multichannel system

The binaural localization theory is utilized in stereo system, i.e. time and level differences between signals from two loudspeakers are utilized in sound reproduction in order to reconstruct the spatial perception of the audience.

Traditional stereo cannot provide the sense of encirclement and immersion because the perception of the sound environment mainly relies on the lateral reflected sound. Surround sound, which constitutes an extension of stereophony, provides full spatial immersion by using reverberation and reflection. The most typical multichannel surround systems are the Dolby surround system, DTS Digital Surround.

Since loudspeakers in Dolby 5.1 are arranged in the same horizontal plane, the reproduction sound image cannot be extended to three dimensions. In 2009, Dolby laboratory presented ProLogic IIz, which extended Dolby 7.1 with height channels (7.1+2). By reproducing early and late reflections and reverberation, ProLogic IIz provide a much wider range of spatial sound effects such as spatial depth and spatial impression [15]. The ProLogic IIz configuration is showed in Figure 1. Audyssey Dynamic Surround Expansion (DSX) is a scalable technology that expands auditory perception by adding height channels, which is in a similar way to Dolby 9.1.

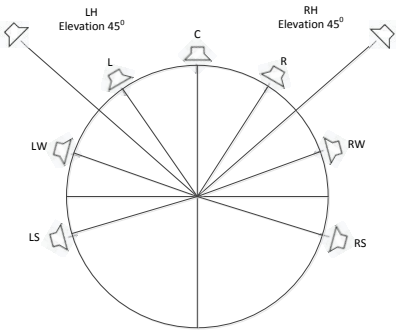


Fig. 1. Dolby IIz configuration

NHK laboratory developed the 22.2 multichannel prototype system in 2003. The system consists of three layers of loudspeakers and overcome the lack of height perception with 3D immersion and sound image localization. K. Hiyama and Keiichi Kubota evaluated the minimum number of loudspeakers and its arrangement for reproducing the spatial impression of diffuse sound field respectively[16]. The results showed that if the interval between adjacent loudspeakers is 45° in both horizontal and vertical plane, there is sufficient horizontal sound envelopment and a good sense of spatial impression. Therefore, the 22.2 multichannel system consists of loudspeakers with a middle layer of ten channels, an upper layer of nine channels, and a lower layer of three regular channels and two Low Frequency Effects (LFE) channels. Figure 2 shows detailed arrangement of loudspeakers[17]. The vertical loudspeaker interval of the 22.2 multichannel is around 45°, which can induce the vertical spatial uniformity [18]. The 22.2 multichannel system reproduces sound images in all three dimensional directions around a listener and stable sound localization over the entire screen area. Subjective evaluations shows that subjects have better impressions using

Ultrahigh-Definition TV (UHDTV) contents with 22.2 multichannel sound system than with Dolby5.1 system[19].

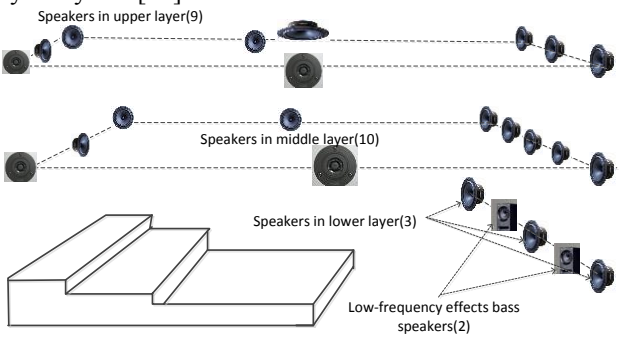


Fig. 2. 22.2 multichannel system layout

2.4 Problems of existing 3D audio systems

Not need to know the loudspeaker layout at the encoding stage is the main advantage of Ambisonics, at decoding stage the loudspeaker signal can be counted according to the loudspeaker layout. The encoding format is an effective reconstruction of 3D sound field, allowing for direct dealing with the three dimensional space characteristics of the sound field such as rotation and mirroring. But along with the increase of order, more precise direction information is carried by spherical harmonic functions, which provides a more accurate location. But data quantity increases rapidly, which requires higher CPU processing power. In addition, the hypothesis that the location of the listener is known may lead to a limit listening area.

The character of WFS is that Kirchhoff-Helmholtz integral can ensure the rebuilt sound field synthesized by secondary sources is the same as the primary source, preserving time and space characteristics of primary source. So listeners can receive and locate the sound source as if it were a real listening space, and walk in the listening area at will while sound image remains unchanged. But WFS needs more loudspeakers and has a higher requirement for site and equipment which is expensive.

The research on compression of Ambisonics and WFS is limited, although recently some progress[20][21][22] has been made. But the compression efficiency cannot meet the requirement of real-time broadcasting and transmitting.

The 22.2 multichannel system, which is based on conventional surround systems plus high and low channels to produce three dimensional sound images, can be easily downmixed for 5.1 system reproduction. It is likely to become a popular 3D system since terminals can be set up with little cost using simplified configuration (10.1 and 8.1 channels), especially when the 5.1 system has already been installed. In 2011, ITU (Report BS.2159-2) pointed out that the 22.2 multichannel system has some problems to be solved: The method to localize more efficiently by using the upper and lower layers and how to reproduce three dimensional sound image movements. In addition, although it is not difficult to downmix 22.2 channel signals to 5.1 channel signals, the 3D spatial audio effects are discarded. Hence, producing three dimensional effects in

home entertainment environments with limited loudspeakers is a problem. Furthermore, without compression, the data rate of 22.2 system can reach 28Mbps and the size of an one-hour audio file is about 100Gb. As a result, it is not possible for the current storage device and transmission channel to adapt to this enormous data. The application and development of 22.2 multichannel systems are constrained by the technology of compression.

3. Hearing mechanism and compression research in 3D audio

3.1 The research of hearing mechanism

From mono, stereo, surround sound, and then to the 3D audio, the main line of development in audio systems is to extend the range of the sound image. Audiences are able to locate the sound which is any position around them in order to bring them a better sense of encirclement and immersion. The positioning of spatial orientation for sound sources is an important content of 3D audio, while the study of perceptual characteristics is an important research field of 3D audio. For example, the arrangement position of the 24 speaker in 22.2-channel system is based on the test and analysis of the angle resolution of sound in horizontal and vertical plane by human ear. In addition, the perceptual research of spatial orientation parameters for sound source is also important for the efficient encoding of the multi-channel audio signal. Therefore, the perceptual characteristics of sound source localization parameters in the 3D sound field are an important way to solve the problems of 3D audio systems.

The perceptual sensitivity of the sound source in the horizontal plane is significantly better than that of the vertical plane or distance by the human auditory system. In the horizontal plane, the positioning of the sound source is dependent on the two binaural cues: ITD and ILD. The human ear can perceive a change in sound image orientation only when the difference of binaural cues reaches a certain threshold value. This threshold value is known as Just Noticeable Difference (JND).The influencing factors of JND for binaural cues are various, including frequency and orientation of the sound source. A wide range of measurements and analysis of these factors has been performed.

Hershkowitz in 1969 [23] and Mossop in 1998 [24] have been researching the influence of sound source position on the perceptual threshold JND of ITD and ILD. The results show that the greater the difference of left and right channels in intensity and time, the larger the JND value of the human perception. This shows that the human ear is less sensitive when the sound source is closer to the left and right sides.

Millers in 1960 measured JNDs of ILD on the midline with pure tones and there were 5 Normal-Hearing (NH) subjects took part in the experiment[25]. The result is as follows: JNDs were around 1dB for 1000Hz, around 0.5dB for frequencies higher than 1000Hz and somewhat smaller than 1dB for frequencies lower than1000Hz. The test data showed worse sensitivity of ILD at 1000Hz than at either higher or lower frequencies. Larisa in 2011 has been researching the influence of the frequency of the

signal on the JND of ITD. The results showed that the perceptual threshold of ITD has a strong dependence on the frequency[26].

The measurement data of JND for binaural cues were fragmented and the conclusions were generally described qualitatively for perceptual threshold of binaural cues. It is difficult to perform mathematical analysis and model accurately and cannot fully reveal the principal of the perceptual threshold of binaural cues. So the JND measurement of binaural cues in all-round, full-band and the mathematical analysis are important issues to reveal the perceptual characteristics of binaural cues. In order to solve the above problem, we have undertaken the research of perceptual characteristics for binaural cues:

In order to study the impact of the frequency and direction on binaural cues JND, our team measured full band JND of binaural cues and analyzed its statistics and distribution characteristics.

a. *Subjects*. 12 NH subjects participated in this study, 7 males and 5 females, all subjects were aged between 19 and 25 years.

b. *Stimuli*. The method in this article used a two-alternative-forced-choice paradigm to measure the JND. Both reference and test signals were 250 ms in duration including 10 ms raised-cosine onset and offset ramps. They were randomly combined into stimulus and separated by 500 ms duration. The stimuli were create by personal computer and presented to the subjects over headphones (Sennheiser HDA 215) at a level of 70 dB SPL. In order to exclude other factors influence on this experiment, the environment of the entire testing process should be consistent and the intensity of test sound must remain around 70 dB SPL. Meanwhile the ITD should be zero in the whole experiment in order to remove the effect on the result caused by other binaural cues and the sum of energy of left and right channels should remain unchanged.

The reference values of ILD in these experiments were 0, 1, 3, 5, 8 and 12 dB, which respond to 6 azimuths(about 0~60 °) in the horizontal plane from midline to the direction of the left ear.

The whole frequency domain was divided into 20 sub-bands, each frequency sub-band satisfied the same perceptual characteristics of human ear.

The stimuli are pure tones whose frequencies are the center frequencies of sub-bands, these frequencies are 75, 150, 225, 300, 450, 600, 750, 900, 1200, 1500, 1800, 2100, 2400, 2700, 3300, 4200, 5400, 6900, 10500, 15500Hz.

c. *Method*. Discrimination thresholds were estimated with an adaptive procedure. During any given trial, subjects would listen to two stimuli by activating a button on a computer screen by mouse-click, with a free number of repeats but the order of two stimulus were changed. The subjects should indicate which one was lateralized to the left relatively by means of an appropriate radio button response in 1.5 s.

An adaptive, 1-up-3-down method was also used in this article. The difference of ILD in dB was increased in every one wrong or decreased in every three consecutive correct judgments. The difference between reference and test signals in first trials was the initial variable which was much larger than the target JND, it was changed by an given step according to previous test results.

The step was changed adaptively, it was adjusted by 50% for the first two reversals, 30% for the next two reversals, then linear changed in a small step size for

the next three reversals, until the final step size reach the expected accuracy for the last three reversals. In a transformed-up-down experiment, the stimulus variable and its direction of change depends on the subjects responses. The direction alternates back and forth between “down” and “up”. Every transform between “down” and “up” was defined as a reversal.

Because of heavy workload of these experiments, adaptive test software was designed to simplify the experiments and the process of data collection and analysis. The software automatically generated test sequences and played one after another. According to the listener’s choice, the software changed ILD values of test stimulus properly, and saved the results to excel sheet until listener hardly distinguished the orientation differences between two sequences. And the value of ILD at this time was the JND value.

d. *Results*. After a subjective listening test for half a year, we got 120 groups(six azimuths and twenty frequencies) of data, each group containing 12 JNDs corresponding to 12 subjects. For every group, we select the data that has the confidence degree of 75% to be JND in that condition. Some JND curves in different reference of ILD were plotted in figure 3:

•The curves vary with the reference ILD, the larger the reference ILD, the higher the corresponding curve. The JND is the most sensitive in the central plane for human perception, and the least sensitive at lateral.

•Human ear is most sensitive to the middle frequency bands except 1000 Hz and less sensitive to the high frequency bands and low frequency bands.

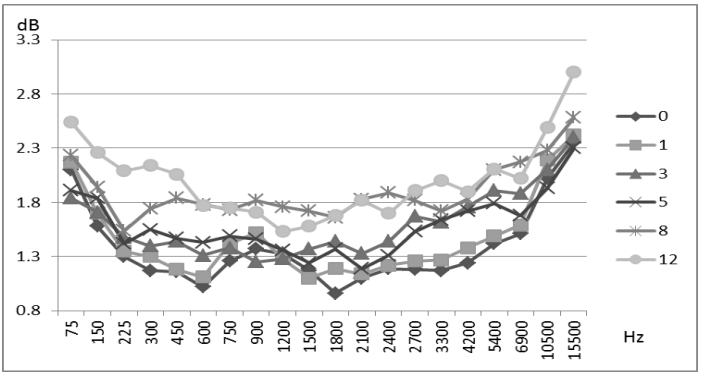


Fig. 3. JND curve of ILD

A binaural perceptual model is established and used in quantisation of ILD. It solves the problem of the perceptual redundancy removal of spatial parameters. Experimental results show that this method can reduce the bitrate by about 15% compared with parametric stereo, while maintaining the subjective sound quality.

3.2 Perceptual information measurement for multichannel audio signal

Multimedia contents abound with subjective irrelevancy—objective information we cannot sense. For audio signals, this means lossless to the extent that the distortion after decompression is imperceptible to normal human ears (usually called transparent coding). The bitrate can be much lower than for true lossless coding. Perceptual audio coding [27] by removing the irrelevancy greatly reduces communication bandwidth or storage space. Psychoacoustics provides a quantitative theory on this irrelevancy: the limits of auditory perception, such as the audible frequency range (20–20000 Hz), the Absolute Threshold of Hearing (ATH), and masking effect[28]. In state-of-the-art perceptual audio coders, such as MPEG-2/4 Advanced Audio Coding (AAC), 64 kbps is enough for transparent coding[29]. The Shannon entropy cannot measure the perceptible information or give the bitrate bound in this case.

For perceptual audio coding technology, determining the lower limit bitrate for transparent audio coding is an important question. Perceptual Entropy (PE) gives an answer to this question[30], which shows that a large amount of audio with CD quality can be compressed with 2.1 bit per sample. PE indicates the least number of bits for quantising mono audio channel without perceptual distortion. This is widely used in the design of quantisers and fast bit allocation algorithm.

Nevertheless, PE has significant limitations when measuring perceptual information. This limitation primarily comes from the underlying monaural hearing model. Humans have two ears to receive sound waves in a 3D space: not only is the time and frequency information perceived— needing just individual ears—but also spatial information or localization information—needing both ears for spatial sampling. Due to the unawareness of binaural hearing, PE of multichannel audio signals is simplified to the supposition of PE of individual channels. This is significantly larger than real quantity of information received because multichannel audio signals usually correlate.

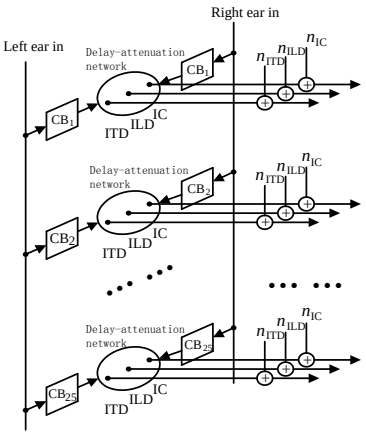


Fig. 4. Binaural Cue Physiological PerceptionModel (BCPPM).

Following the concept of PE, we establish a Binaural Cue Physiological Processing Model (BCPPM, figure 4). Based on MCPPM, we using EBR filter to simulate the human cochlea filter effect, and the JND of binaural cues to estimate the absolute threshold of spatial cues.

a. *SPE Definition*. From the information theory viewpoint, we see BCPPM as a double-in-multiple-out system (Figure 4). The double-in is the left ear entrance sound and the right ear entrance sound. The multiple-out consists of 75 effective ITDs, ILDs, and ICs (25 CBs, each with a tuple of ITD, ILD, and IC). Like in computing PE, we view each path that leads to an output as a lossy subchannel. Then there are 75 such subchannels. Unlike PE, what a subchannel conveys is not a subband spectrum but one of ITD, ILD, and IC of the subband corresponding to the sub-channel. In each sub-channel, there are intrinsic channel noises (resolution of spatial hearing), and among sub-channels, there are interchannel interferences (interaction of binaural cues). Then there is an effective noise for each sub-channel. Under this setting, each sub-channel will have a channel capacity. We denote SPE(c), SPE(t), and SPE(l) for the capacity of IC, ITD, and ILD sub-channels respectively. Then SPE is defined as the overall capacity of these sub-channels, or the sum of capacities of all the sub-channels:

$$SPE = \sum_{all\ subbands} SPE(c) + SPE(t) + SPE(l) \tag{1}$$

To derive SPE(c), SPE(t), and SPE(l), we need probability models for IC, ITD, and ILD. Although the binaural cues are continuous, the effective noise quantizes them into discrete values. Let [L • P], [T • P], and [C • P] denote the discrete ILD, ITD, and IC source probability spaces:

$$\begin{aligned}
[\mathbf{L}\bullet P]&:\begin{cases} \mathbf{L}: & l_1, & l_2, & \cdots, & l_i, & \cdots, & l_N \\ P(\mathbf{L}): & P(l_1), & P(l_2), & \cdots, & P(l_i), & \cdots, & P(l_N) \end{cases} \\
[\mathbf{T}\bullet P]&:\begin{cases} \mathbf{T}: & t_1, & t_2, & \cdots, & t_i, & \cdots, & t_N \\ P(\mathbf{T}): & P(t_1), & P(t_2), & \cdots, & P(t_i), & \cdots, & P(t_N) \end{cases} \\
[\mathbf{C}\bullet P]&:\begin{cases} \mathbf{C}: & c_1, & c_2, & \cdots, & c_i, & \cdots, & c_N \\ P(\mathbf{C}): & P(c_1), & P(c_2), & \cdots, & P(c_i), & \cdots, & P(c_N) \end{cases}
\end{aligned} \tag{2}$$

where l_i, t_i, and c_i are the ith discrete values of ILD, ITD, and IC, respectively, and $P(l_i)$, $P(t_i)$, and $P(c_i)$ the corresponding probabilities. Then we have

$$\begin{aligned}
SPE(l) &= -\sum_{i=1}^{N_L} p(l_i)\log_2 p(l_i) \\
SPE(t) &= -\sum_{i=1}^{N_T} p(t_i)\log_2 p(t_i) \\
SPE(c) &= -\sum_{i=1}^{N_C} p(c_i)\log_2 p(c_i)
\end{aligned} \tag{3}$$

b. *CB Filterbank*. We use the same method as that in PE to implement the CB filterbank. Audio signals are first transformed to the frequency domain by DFT of 2048 points with 50% overlap between adjacent transform blocks. Then a DFT spectrum is partitioned into 25 CBs.

c. *Binaural Cues Computation*. ILD, ITD, IC are computed in the DFT domain as described in [31].

d. *Effective Spatial Perception Data* . The resolutions or quantization steps of the binaural cues can be determined by JND experiments. Denote by $\Delta\tau$, $\Delta\lambda$, and $\Delta\eta$ the

resolutions of ITD, ILD, and IC, respectively. Generally, they are signal dependent and frequency dependent. For simplicity, we use constant values: $\Delta\tau = 0.02$ ms, $\Delta\lambda =$ 1dB, and $\Delta\eta = 0.1$.

We ignore the effect of IC on ILD and only consider the effect of IC on ITD for SPE computation. Lower IC leads to lower resolution of ITD. This is equivalent to higher JND of ITD. Then the effective JND on subband b, denoted as $\Delta\tau'$ (b), can be formulated as the following:

$$\Delta\tau'(b) = \frac{\Delta\tau(b)}{\mathrm{IC}(b)} \tag{4}$$

Then we have the following effective perception data q_{ILD}(b), q_{ITD}(b), and q_{IC}(b) of ILD, ITD, and IC, respectively by quantization, where $\lfloor \cdot \rfloor$ represents the round down function:

$$\begin{aligned} q_{\mathrm{ILD}}(b) &= 2\left\lfloor \left| \frac{\mathrm{ILD}(b)}{\Delta\lambda(b)} \right| \right\rfloor \\ q_{\mathrm{ITD}}(b) &= 2\left\lfloor \left| \frac{\mathrm{ITD}(b)}{\Delta\tau(b)/\mathrm{IC}(b)} \right| \right\rfloor \\ q_{\mathrm{IC}}(b) &= \left\lfloor \frac{1\text{-}\mathrm{IC}(b)}{\Delta\eta(b)} \right\rfloor \end{aligned} \tag{5}$$

Suppose that $q_{\mathrm{ILD}}(b)$, $q_{\mathrm{ITD}}(b)$, and $q_{\mathrm{IC}}(b)$ are uniformly distributed by (3), the SPE are expressed as

$$\begin{aligned} SPE = \frac{1}{N}\sum_{b=1}^{25} & \alpha \log_2\left(\mathrm{int}\left(\frac{1-\mathrm{IC}(b)}{\Delta\eta(b)} \right) + 1 \right) + \alpha \log_2\left(2\,\mathrm{int}\left(\left| \frac{\mathrm{ITD}(b)}{\Delta\tau(b)/\mathrm{IC}(b)} \right| \right) + 1 \right) \\ & + \alpha \log_2\left(2\,\mathrm{int}\left(\left| \frac{\mathrm{ILD}(b)}{\Delta\lambda(b)} \right| \right) + 1 \right) \end{aligned} \tag{6}$$

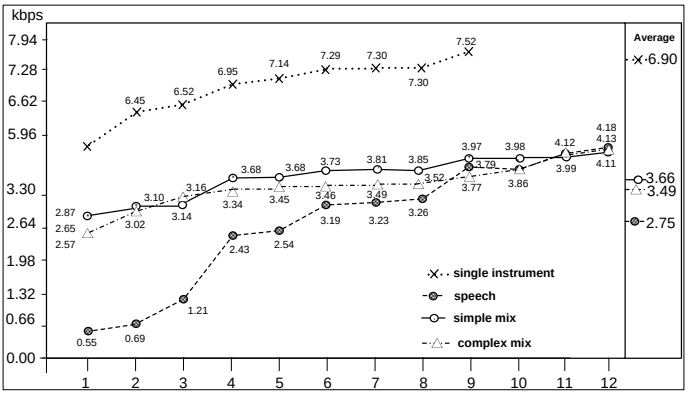


Fig. 5. Perceptual spatial information of stereo sequences sampled at 44.1 kHz.

d. *Results*. Figure 5 shows the SPE of four different stereo signals from MPEG test sequences. The experiment suggests that SPE of speech signal is very low. This is because the human voice is often recorded with fixed position without change. So coding this kind of stereo audio signals requires a low bit rate. The average SPE for

speech signals is 2.75kbps, for simple mixed audio is 3.66kbps, for complex mixed audio is 3.49kbps and for a single instrument is 6.90kbps. In other words, to achieve transparent stereo effect, audio signals required more than 7kbps, which is close to the bitrate 7.7kbps of PS. So the proposed SPE can reflect the amount of perceptual spatial information that is ignored by PE. Experiments on stereo signals of different types have confirmed that SPE is compatible with the spatial parameter bitrate of PS.

Using PE to evaluate the perceptual information, only interchannel redundancy and irrelevancy are exploited; the overall PE is simply the sum of PE of the left and right channels. Using SPE based on BCPPM, interchannel redundancy and irrelevancy are also exploited; the overall perceptual information is about one normal audio channel plus some spatial parameters, which has significantly lower bitrate. For the above reason, PE gives much higher bitrate bound than SPE. PE is compatible with the traditional perceptual coding schemes, such as MP3 and AAC, in which channels are basically processed individually (except the mid/side stereo and the intensity stereo). So PE gives meaningful bitrate bound for them. But in Spatial Audio Coding (SAC), multichannel audio signals are processed as one or two core channels plus spatial parameters. SPE is necessary in this case and generally gives much lower bitrate bound (~1/2). This agrees to the sharp bitrate reduction of SAC.

4. Tendency of 3D audio technology and our future work

4.1 Hearing mechanism research on 3D audio

The spatial orientation cues of sound include three aspects: azimuth angle, elevation angle and distance. There are many acoustic factors to perceive the distance of a sound source, such as the source of the sound (sound pressure level and spectrum), the transmission environment (reflected sound, high-frequency losses and environmental noise) as well as listening factors. So the current research focuses on the expression and extraction of distance cues. Hence, the perceptual characteristic of the 3D spatial orientation is an important research direction for 3D audio technology.

Our future work will focus on the perceptual characteristics of 3D spatial orientation. The main work will include: design experiments to obtain perceptual threshold of 3D spatial position, mathematical analysis to establish representation model of perceptual sensitivity in 3D spatial orientation, get the perceptual distortion of sound image in the different offset of spatial orientation, obtain the equivalent distortion curve of azimuth angle and elevation angle in 3D spatial orientation, and to establish a position distortion model of 3D spatial position. Through the above research, we expect to establish the basic theory of perceptual mechanism for 3D audio systems and provide theoretical support for 3D audio collection, processing, reconstruction, playback and evaluation.

4.2 High efficiency compression for 3D audio signal

Existing 3D audio compression technology has exploited the perceptual redundancy within each individual channel. From the same sound field and same sound source, 3D audio signals of different channels intrinsically exhibit strong correlation. Parametric coding is able to extract the cues of sound image direction, width and scene information to reduce the interchannel redundancy, and achieve high compression efficiency using fewer channels with side information. Parametric coding for 3D audio is able to fulfill the compression requirement of transmission and storage while keep 3D effect meantime, so it is a strong direction in 3D audio compression research.

Since the compression is highly efficient, the reconstructed 3D effect strongly depends on the cues that described corresponding spatial information. The existing 3D audio parameter coding quantises those cues uniformly and reconstruction error in every direction is the same. However, according to human perceptual characteristic in 3D space, the JND to sound direction exists and varies widely in all directions. If reconstruction error for direction cues exceed corresponding threshold, perceptible 3D effect distortion is produced. So how to utilize human perceptual characteristics in 3D space for 3D audio parametric coding will be included in our future work. Our goal is to develop the 3D spatial perception information measurement and establish a computational model of 3D audio orientation perception for effective representation of 3D audio parameterization

4.3 The evaluation of 3D audio quality

Along with the developments of the 3D audio technology, research institutions such as NHK [32] and Deutsche Telekom Laboratories[33], are carrying out the subjective evaluation of the 3D audio system. Because the subjective evaluation is based on the human who is the main body directly involved in the evaluation, the result is more explicit and reasonable in spite of spending a lot of time and manpower during the period of the assessments. So, more and more scholars[34][35][36] are trying to establish the objective evaluation model for the 3D audio system, hoping to look for an objective evaluation model based on the human perception of the audio quality to assess the effects of a 3D sound field. The performance of the proposed model is comparable with the subjective evaluation method.

However, the current methods used to establish an objective evaluation model do not introduce the spectral cues related to the elevation perception of sound events, the envelopment or immersion in diffuse sounds, or the proximity and distance of sound events as the acoustic characteristic parameters. Research of the objective evaluation methods of the 3D audio is occuring on to investigate the spectral cues of the elevation, envelopment and distance perception of the 3D sound field.

In the study of the objective evaluation method of the 3D audio quality, we draw up an objective evaluation model, based on the acoustic characteristic parameters of a 3D audio signal, to predict the perceptual acoustic attributes of the 3D sound field. Including the Basic Audio Quality (BAQ), the Timbral Fidelity (TF), the 3D Frontal Spatial Fidelity (3DFSF) and the 3D Surround Spatial Fidelity (3DSSF). The study includes establishing the acoustic characteristic parameter set related to the 3D

perceptual sound field, obtaining a predictable mapping of the perceptual acoustic attributes and the acoustic characteristic parameters of a 3D audio quality, and building up an objective evaluation model of the 3D perceptual sound field by fitting the performances of the subjective evaluation and objective evaluation. Because the main aim of this study is to express the spectral cues related to the elevation perception of a 3D sound field, we should try to analyze the duplex spectral effects of the pinna to further improve the technology of the 3D audio objective evaluation.

5. Conclusion

The complexity and large capacity limit the promotion and application of 3D audio. To solve these problems, the National Natural Science Foundation of China, Tsinghua University, Wuhan University and other colleges organized the Second International Symposium of 3D video and audio. In the 3D audio workshop, basic theory and research on the recording, compression and reconstruction for 3D audio was emphasized. We also hope to promote the research work to become part of the next generation standard for the audio and video coding (AVS2) of China. This paper gives a brief introduction on current 3D audio systems. At the same time, our research work on the hearing mechanism and compression coding are presented. Finally our future work is introduced, which includes the research of perception characteristic, compression coding and the quality evaluation.

Acknowledgment

This work is supported by National Natural Science Foundation of China (No.60832002, No.61102127), Major national science and technology special projects (2010ZX03004-003-03), Nature Science Foundation of Hubei Province (2010CDB08602, 2011CDB451), Wuhan ChenGuang Science and Technology Plan (201150431104), and the Fundamental Research Funds for the Central Universities.

References

1. Berkhout, A.J.: A holographic approach to acoustic control. Journal of the Audio Engineering Society. 36, 977-995 (1988)
2. Berkhout, A., De Vries, D., Vogel, P.: Acoustic control by wave field synthesis. J. Acoust. Soc. Am. 93, 2764-2778 (1993)
3. R. Rabenstein, S.S., P. Steffen: Wave field synthesis techniques for spatial sound reproduction. In: Topics in Acoustic Echo and Noise Control, pp. 517-545. Springer, Berlin, Heidelberg (2006)
4. De Vries, D.: Wave Field Synthesis: History, State-of-the-Art and Future (Invited Paper). In: Universal Communication, 2008. ISUC '08. Second International Symposium on, pp. 31-35.

(2008)
5. De Bruijn, W.:Application of wave field synthesis in videoconferencing. Delft University of Technology (2004)
6. Vogel, P.:Application of Wave Field Synthesis in Room Acoustics. Delft University of Technology (1993)
7. Daniel, J.M., Sebastien; Nicol, Rozenn: Further Investigations of High-Order Ambisonics and Wavefield Synthesis for Holophonic Sound Imaging. In: Audio Engineering Society Convention 114. Amsterdam, The Netherlands (2003)
8. Gerzon, M.A.: Ambisonics: Part two: Studio techniques. Studio Sound. (1975)
9. Malham, D.G.: Spatial hearing mechanisms and sound reproduction. University of York. (1998)
10. Furness, R.K.: Ambisonics-an overview. In: 8th International Conference: The Sound of Audio, pp. 181-189. (1990)
11. Keating, D.: The generation of virtual acoustic environments for blind people. In: Proc.1st Euro. Conf: Disability, Virtual Reality & Assoc. Tech., pp. 201-207. Maidenhead, UK (1996)
12. Elen, R.: Whatever happened to Ambisonics? AudioMedia Magazine, November. (1991)
13. Gerzon, M.A.: Ambisonics in multichannel broadcasting and video. J. Audio Eng. Soc. 33, 859-871 (1985)
14. Strutt, J.W.: On our perception of sound direction. Philosophical Magazine. 13, 214-232 (1907)
15. Theile, G., Wittek, H.: Principles in Surround Recordings with Height. In: Audio Engineering Society Convention 130. (2011)
16. Hiyama, K., Komiyama, S., Hamasaki, K.: The minimum number of loudspeakers and its arrangement for reproducing the spatial impression of diffuse sound field. Audio Engineering Society Convention 113. (2002)
17. Ando, A.: Home Reproduction of 22.2 Multichannel Sound. In: 5th International Universal Communication Symposium. (2011)
18. Oode, S., Sawaya, I., Ando, A., Hamasaki, K., Ozawa, K.: Vertical Loudspeaker Arrangement for Reproducing Spatially Uniform Sound. In: Audio Engineering Society Convention 131. (2011)
19. Hamasaki, K., Nishiguchi, T., Okumura, R., Nakayama, Y., Ando, A.: A 22.2 multichannel sound system for ultrahigh-definition TV (UHDTV). Smpte Motion Imaging Journal. 117, 40-49 (2008)
20. Cheng, B., Ritz, C., Burnett, I.: A Spatial Squeezing approach to Ambisonic audio compression. In: IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP), pp. 369-372. IEEE, (2008)
21. Hellerud, E., Solvang, A., Svensson, U.P.: Spatial redundancy in Higher Order Ambisonics and its use for lowdelay lossless compression. In: Acoustics, Speech and Signal Processing, 2009(ICASSP 2009). IEEE International Conference on, pp. 269-272. (2009)
22. Pinto, F., Vetterli, M.: Space-Time-Frequency Processing of Acoustic Wave Fields: Theory, Algorithms, and Applications. Signal Processing, IEEE Transactions on. 58, 4608-4620 (2010)
23. Hershkowitz, R., Durlach, N.: Interaural Time and Amplitude JNDs for a 500 - Hz Tone. The Journal of the Acoustical Society of America. 46, 1464-1465 (1969)
24. Mossop, J.E., Culling, J.F.: Lateralization of large interaural delays. The Journal of the

Acoustical Society of America. 104, 1574-1579 (1998)
25. Mills, A.W.: Lateralization of High‐Frequency Tones. The Journal of the Acoustical Society of America. 32, 132-134 (1960)
26. Dunai, L., Hartmann, W.M.: Frequency dependence of the interaural time difference thresholds in human listeners. The Journal of the Acoustical Society of America. 129, 2485-2485 (2011)
27. Painter, T., Spanias, A.: Perceptual coding of digital audio. Proceedings of the IEEE. 88, 451-515 (2000)
28. Moore, B.C.J.: Masking in the Human Auditory System. In: Audio Engineering Society Conference: Collected Papers on Digital Audio Bit-Rate Reduction. Audio Engineering Society, New York, USA (1996)
29. Bosi, M., Goldberg, R.E.: Introduction to digital audio coding and standards. Kluwer Academic Publishers, Boston, Mass, USA (2003)
30. Johnston, J.D.: Transform coding of audio signals using perceptual noise criteria. Selected Areas in Communications, IEEE Journal on. 6, 314-323 (1988)
31. C. Faller, F. Baumgarte. :Binaural cue coding—part II: schemes and applications. IEEE Transactions on Speech and Audio Processing, vol. 11, no. 6, pp. 520–531 (2003)
32. Hamasaki, K.H., Koichiro; Nishiguchi, Toshiyuki; Okumura, Reiko: Effectiveness of Height Information for Reproducing the Presence and Reality in Multichannel Audio System. In: Audio Engineering Society Convention 120. Paris, France (2006)
33. Geier, M., Wierstorf, H., Ahrens, J., Wechsung, I., Raake, A., Spors, S.: Perceptual evaluation of focused sources in wave field synthesis. In: AES 128th Convention, pp. 22-25. (2010)
34. George, S.:Objective models for predicting selected multichannel audio quality attributes. Department of Music and Sound Recording, University of Surrey (2009)
35. Epain, N., Guillon, P., Kan, A., Kosobrodov, R., Sun, D., Jin, C., Van Schaik, A.: Objective evaluation of a three-dimensional sound field reproduction system. In: Proceedings of 20th International Congress on Acoustics. Sydney, Australia (2010)
36. Song, W., Ellermeier, W., Hald, J.: Psychoacoustic evaluation of multichannel reproduced sounds using binaural synthesis and spherical beamforming. The Journal of the Acoustical Society of America. 130, 2063-2075 (2011)

Rolling Sound Synthesis : Work In Progress

Simon CONAN, Mitsuko ARAMAKI, Richard KRONLAND-MARTINET and Sølvi YSTAD*

Laboratoire de Mécanique et d'Acoustique, MARSEILLE, FRANCE
{conan , aramaki , kronland , ystad}@lma.cnrs-mrs.fr

Abstract. This paper presents a physically informed rolling sound synthesis model for the *MétaSon* synthesis platform. The aim of this sound synthesis platform will be shortly described. As shown in the state of the art, both in terms of sound effects and proposed controls, existing models can be improved. Some details on asymmetric rolling objects will be given and the sound synthesis model will be exposed. Perspectives for further studies and work in progress will be discussed.

Keywords: Rolling Sounds, Sound Synthesis and Control, Sound Invariants, Physically Informed Synthesis, Rolling ball, Environmental Sounds Synthesis

1 Introduction

This study is part of a larger project (*MétaSon*) whose aim is to build a real-time sound synthesis platform that offers a perceptual control of sounds. To this purpose we need to :

- Build real-time synthesis models that reproduce sound features of real objects (plates, shell, water...) and interactions between them (rolling, rubbing...).
- Associate perceptual control strategies to these sound synthesis models control. By perceptual control, we mean that the synthesis model should be controllable by words which describe the sound as the user perceives it (e.g. "I want to produce a sound which rolls and with a liquid texture" or "I want a wind that sounds metallic") or by gestures.
- Identify sound *invariants* in order to achieve such perceptual controls. These *invariants* are either *structural invariants*, i.e. sound morphologies responsible for the recognition of an object and its properties, or *transformational invariants*, i.e. sound morphologies responsible for the recognition of the action on the object (see for example [1] for a study on breaking and bouncing invariants, and [2–4] for a more general approach). We are convinced that these *invariants* are strong enough to evoke both actions and objects and that it is possible to build and control sounds from perceptual categorizations in order to construct sound metaphors, like "bouncing water" or "rolling wind" for example.

* The authors would like to thank the French National Research Agency (ANR) which supports the *MetaSon* Project - CONTINT 2010 : ANR-10-CORD-010.

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

In this paper, we propose to examine rolling sound synthesis and possible control strategies associated to such sounds. Different approaches to the synthesis of rolling sounds can be found in the literature.

One approach is the physical modeling of the phenomenon and the computation of equations with finite difference scheme. Stoelinga et al. derived a physical model that produce rolling sounds [5] from previous studies on impact sounds on damped plates [6, 7]. This model can reproduce effects like Doppler which is also found in the measures. However, sound examples aren't fully convincing, i.e. the sounds don't evoke rolling balls without ambiguity. This can be explained by the fact that no amplitude modulation is present in the case of continuous contact as the rolling object is considered as a perfect sphere (i.e. the mass center is the geometrical center). They also simulated very special cases of rolling like periodic bouncing and their simulations are comparable to their measures. It is important to note that these models cannot be computed in real time.

Another approach is the physically informed modelling. Here, the aim is to reproduce the "sound effect" produced by a rolling object. The sensation of a rolling object can then be modified by acting on specific acoustic features. These models are generally source-filter synthesis models, i.e. a source excitation which passes through a filter-bank (resonant object), informed by phenomenological considerations. The user can hereby control the synthesis parameters to act on size and speed of the rolling balls for example, but the control of these parameters can sometimes be difficult because the synthesis model has been constructed empirically. Van den Doel et al. [8] proposed a model where modal resonators were fed with a noise whose spectral envelope was defined by $\sqrt{1/(\omega-\rho)^2+d^2}$ where ρ and d are respectively the frequency and the damping of the resonance, in order to enhance the resonance near the rolling object's mode.

In order to extract parameters from real recorded sounds for a sound synthesis model (i.e. parameters of a source-filter model), Lagrange et al. [9] and Lee et al. [10] proposed an analysis/synthesis scheme. The aim was to extract the excitation pattern and the object's resonances (the resonance of the rolling object and the surface on which it rolls were not separated). Depending on the recording that is analyzed, this can yield good resynthesis, but no general model of rolling objects with associated controls can be derived from such methods.

Both in terms of sound effects and proposed controls, the previously presented models can be improved. In fact, if we examine sound features that seem important for the perception of rolling sounds, we can conclude that more cues are necessary to perceive a wide variety of object sizes and rolling speeds. Houben et al. studied the auditory ability to distinguish the largest or the fastest ball between two recorded sounds. They also attempted to identify acoustic cues that characterize the size and speed of rolling balls, like auditory roughness or spectral structure that could be used to identify size and speed of rolling balls. They showed that at constant velocity (respectively at constant size) listeners can distinguish the largest (respectively the fastest) rolling ball with good results. Performance is impaired when the two factors (i.e. velocity and size) are crossed [11]. The influence of spectral and temporal properties was studied in [12] by

crossing the temporal content of a stimulus with the spectral content of another stimulus and using the obtained sound (the obtained stimulus has its spectrum very close to one stimulus and its temporal envelope very close to the other stimulus) in a perceptual experiment. It is shown that only the spectral structure is used to determine the fastest or largest ball and that results are better for the size judgement than for the speed judgement. However only recordings without clear amplitude modulation (due to an unbalanced ball or a deviation from perfect sphericity) were used in the experiment. This can explain why no temporal cues were found. The influence of this amplitude modulation is addressed in [13]. Artificial amplitude modulations were added to the recordings used in the previous experiments. Perceptual experiments showed that amplitude modulations clearly influence the perceived size and speed.

We would like to improve the synthesis of rolling by including the modulation effects. This problem was already addressed by some authors. In [14], Hermes proposed a synthesis model. It consisted of a series of impacts following a Poisson law amplitude modulated to account for the asymmetry of the ball. This pattern was further convolved with a sum of gamma-tones to represent the impulse response of the object on which the ball rolled. He justified this form of impulse response in comparison to the classical representation that uses a sum of exponentially decaying sinusoids by the fact that the collisions between the ball and the plate are "softer". The control of this model is quite complicated. In [15], Rath described a physically informed rolling sound synthesis model. Impacts on the surface were modeled by modal resonators. These resonators were fed with a low-pass filtered noise (which represents the surface profile) which was further filtered by the "rolling filter" to simulate the irregularities encountered by the ball. The force was modulated by a sinusoid to account for asymmetry of the rolling ball and the whole model could be run in real-time. Nevertheless, the modulation force was only derived for a constant velocity.

In the rest of this paper, we will describe the rolling sound synthesis model we developed. First, a simplified model of amplitude modulation due to an unbalanced rolling ball will be exposed and then the global synthesis model will be presented. In the last section, the work in progress and the perspectives will be presented.

2 Asymmetric rolling object

The aim of this section is to get an idea of the modulation profile generated by an imperfect rolling ball (i.e. a ball for which the mass center differs from the geometrical center). Our model is very simple and does not reflect the real motion of the ball because we impose a given velocity profile of the ball's geometrical center to get the modulation profile (in fact, the mass center's eccentricity impose a velocity to the geometrical center). We use this approximation because the kinematics of an unbalanced ball is not a trivial problem and we cannot easily derive solutions from studies on unbalanced rolling objects (see the studies on loaded hoops in [16] for example).

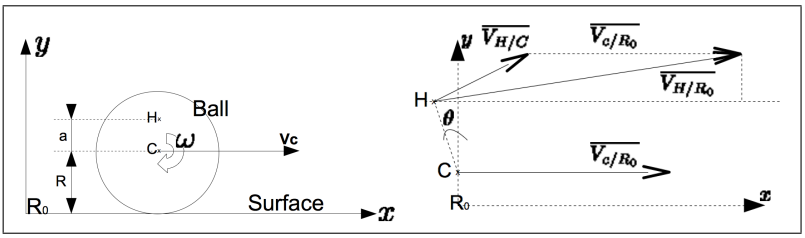


Fig. 1. Left : Representation of a ball of radius R rolling with a transversal velocity V_c. Its mass center H is off centered by a distance a from its geometrical center C. Right : Representation of the velocities.

So let us consider a ball with C its geometrical center and R its radius that rolls with a transversal velocity V_c (angular velocity ω) in the reference frame R_0. The mass center of this ball is situated at the point H, at a distance a from C (see figure 1). Assuming a pure rolling motion, the movement of the point H toward y is :

$$h_y(x) = R + a\sin(2\pi\frac{x}{2\pi R}) = R + a\sin(\frac{x}{R}) \tag{1}$$

The amplitude modulation is due to the height variation of the mass center (which is related to potential energy) with respect to time :

$$h_y(t) = R + a\sin(\frac{x(V_H(t))}{R}) \tag{2}$$

with $V_H(t)$ the velocity of the point H within the frame of reference R_0. The component of $\overline{V_{H/R_0}}$ on the $\overline{x}$ axis (see figure 1 for notations) is (assuming the ball is rolling without sliding : $\|\overline{V_{c/R_0}}\| = R\omega$ and $\|\overline{V_{H/c}}\| = a\omega$) :

$$\overline{V_{H/R_0}} \cdot \overline{x} = \|\overline{V_{c/R_0}}\|(1 + \alpha\cos(\theta)) \tag{3}$$

with $\alpha = a/R$. For more clarity, we will write $\overline{V_{H/R_0}} \cdot \overline{x}$ and $\|\overline{V_{c/R_0}}\|$ respectively V_H et V_c. And θ follows :

$$\theta(t) = \int \omega(t)dt = \int \frac{V_c(t)}{R}dt \tag{4}$$

We can see the modulations for two different asymmetries in figure 2.

3 Sound synthesis model

A general framework of the model is shown in figure 3 with its associated controls. The sound synthesis model is based on a source-filter architecture : a noise (the source) is sculpted by successive filters, then is modulated in amplitude and finally feeds a bank of resonant filters that describe the surface on which the object rolls. The control parameters are based both on perceptual attributes

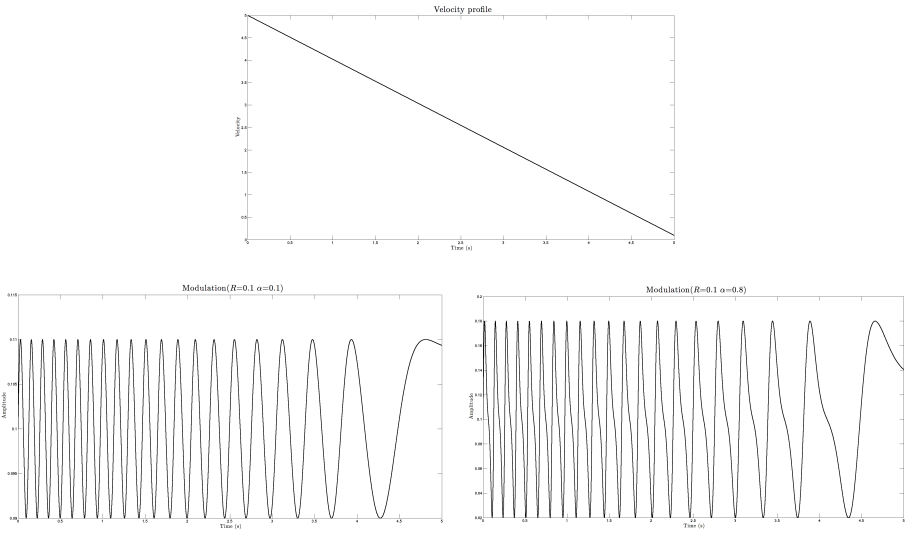


Fig. 2. Velocity profile and associated modulation for two different asymmetries.

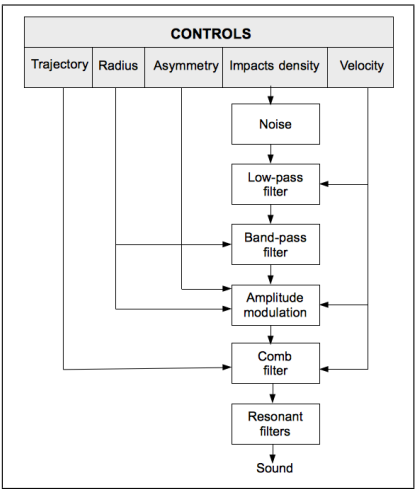


Fig. 3. General Framework of the synthesis model to produce rolling sounds.

and on physical considerations. The links between these controls and the low-level synthesis parameters are described below.

In order to simulate series of collisions between the ball and the surface (i.e. the relative surface which is "seen" by the ball as it is rolling), we use a noise. This signal consists in a sequence of impacts of different amplitudes which are more or less spaced in time. The temporal pattern is modelled as a random process : at each sample, a Bernoulli process is performed with a probability ρ. We choose $\rho \in [0.01, 0.03]$ (a value of ρ that is too high leads to a sound that is too noisy, and a value below 0.01 leads to a sound that is too discontinuous). The amplitude of each impulse is random, and follows an uniform law between 0 and 1. We already used this noise to simulate the sound produced by two continuous interactions, rubbing and scratching [17]. This study focuses on the perceptual differences between rubbing and scratching actions evoked by recorded and synthesized sounds, and shows that a density of $\rho < 0.01$ is associated to scratching and $\rho > 0.1$ is associated to rubbing and that the perception is ambiguous between these two values. This kind of noise is similar to the one used in [14].

This noise is then low-pass filtered. The cut-off frequency is related to the transversal V_c velocity of the ball, i.e. the faster the ball the higher the cut-off frequency (see [8] for further information). To account for the size of the ball, we use a band-pass filter with center frequency $f_c \propto (1/R)$ with R the radius of the ball. The assumption that motivates this filtering stage is that the plate is more excited near the modes of the rolling object.

The amplitude modulation is then applied to the filtered noise. We get the resulted noise s by computing $s^{n+1} = e^{n+1}.h_y^{n+1}$ with e the excitation noise previously described and h_y the height variation of the mass center computed as :

$$h_y^{n+1} = R(1 + \alpha \sin(\frac{x^{n+1}}{R})) \quad \text{with} \quad \begin{cases} x^{n+1} = x^n + V_H^{n+1} dt \\ V_H^{n+1} = V_c^{n+1}(1 + \alpha \cos(\theta^{n+1})) \\ \theta^{n+1} = \theta^n + \frac{V_c^{n+1}}{R} dt \end{cases} \quad (5)$$

To account for the position-dependent excitation, we use a comb filter (see for example the explanation of Smith on the position dependent excitation on a guitar string [18]). This filtering stage is important as it gives the listener a sensation of displacement.

Finally, the obtained excitation is used to feed a resonant filter bank used to simulate the surface on which the ball rolls. These filters model the impulse response of the surface, which is related to the *structural invariant* responsible for the recognition of the surface.

4 Perspectives

This model yields convincing results but the mapping strategy needs to be improved. In fact, the height variation of the center of mass needs to be linked to a

force applied to the surface on which the object rolls. We could derive a force as Rath in [15] by applying Newton's law to the height variation of the mass center $F(t) = M.\ddot{h}_y(t)$, with M the rolling object mass and $\ddot{h}_y(t)$ the acceleration perpendicular to the plain of the mass center. However, the mapping between the obtained force and the synthesis model is not direct and we are currently investigating a mapping between the force and the rolling sounds signal model. The force should be integrated at a lower level in the model, i.e. directly in the random process of micro-impacts generation. We are also working on defining a physical model of rolling objects. Our aim is to try to recover parameters of the model from real recordings thanks to inverse problem methods. Such an approach will validate our force model or show us if further refinements (e.g., reduction of simplifying assumptions) are necessary.

Besides the construction of a synthesis model that simulates realistic rolling sounds, calibrated sounds from this model can also be used to identify perceptual cues responsible for the recognition of this specific action. Furthermore, this approach aimed at highlighting the acoustical cues, also called *invariants*, which characterized the rolling action. Perceptual tests should further validate these assumptions before they are tested as descriptors on real sounds.

Another aspect that would add realism to the rolling sound synthesis is the simulation of the ball's position on the surface. Taking into account the multiple reflections from the edges of the surface on which the object rolls by adding several comb filters for which the delays are computed by a source-image method may increase realism. In [19], Stoelinga et al. analysed the wave dispersion (i.e. the frequency dependent wave velocity) in a plate and concluded that frequency dependent comb filters add more realism when simulating a ball approaching the edge of a plate. Finally, real time implementation should conclude this work.

References

1. Warren, W.H. and Verbrugge, R.R.: Auditory Perception of Breaking and Bouncing Events: A Case Study in Ecological Acoustics. Journal of Experimental Psychology: Human Perception and Performance 10, 5, 704–712 (1984)
2. McAdams, S.E. and Bigand, E.E.: Thinking in Sound: The Cognitive Psychology of Human Audition. Oxford Science Publication. Chapter 6 (1993)
3. Gaver, W.W.: What in the World Do We Hear?: An Ecological Approach to Auditory Event Perception. Ecological psychology, 5, 1, 1–29 (1993)
4. Michaels, C.F. and Carello, C.: Direct Perception. Prentice-Hall Englewood Cliffs, NJ (1981)
5. Stoelinga, C., Chaigne, A.: Time-Domain Modeling and Simulation of Rolling Objects. Acta Acustica united with Acustica 93, 2, 290–304 (2007)
6. Chaigne, A., Lambourg, C.: Time-Domain Simulation of Damped Impacted Plates. I. Theory and Experiments. Journal of the Acoustical Society of America 109 (2001)
7. Lambourg, C., Chaigne, A., Matignon, D.: Time-Domain Simulation of Damped Impacted Plates. II. Numerical Model and Results. Journal of the Acoustical Society of America 109 (2001)
8. Van Den Doel, K., Kry, P.G., Pai, D.K.: FoleyAutomatic: Physically-Based Sound Effects for Interactive Simulation and Animation. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, 537–544 (2001)

9. Lagrange, M., Scavone, G., Depalle, P.: Analysis/Synthesis of Sounds Generated by Sustained Contact between Rigid Objects. IEEE Transactions on Audio, Speech, and Language Processing, 18, 3 509–518 (2010)
10. Lee, J.S., Depalle, P., Scavone, G.: Analysis/Synthesis of Rolling Sounds Using a Source-Filter Approach. In: 13th Int. Conference on Digital Audio Effects (DAFx-10), Graz, Austria (2010)
11. Houben, M.M.J., Kohlrausch, A., Hermes, D.J.: Perception of the Size and Speed of Rolling Balls by Sound. Speech Communication 43, 4, 331–345 (2004)
12. Houben, M.M.J., Kohlrausch, A., Hermes, D.J.: The Contribution of Spectral and Temporal Information to the Auditory Perception of the Size and Speed of Rolling Balls. Acta Acustica united with Acustica 91, 6, 1007–1015 (2005)
13. Houben, M.: The Sound of Rolling Objects, Perception of Size and Speed. Technische Universiteteit, Eindhoven (2002)
14. Hermes, D.J.: Synthesis of the Sounds Produced by Rolling Balls. In: Internal IPO report no. 1226, IPO, Center for User-System Interaction, Eindhoven, The Netherlands (1998)
15. Rath, M.: An Expressive Real-Time Sound Model of Rolling. In: Proceedings of the 6th "International Conference on Digital Audio Effects" (DAFx-03). Citeseer (2003)
16. Theron, W.F.D.: Analysis of the Rolling Motion of Loaded Hoops (2008)
17. Conan S., Aramaki M., Kronland-Martinet R., Thoret E. and Ystad S.: Perceptual Differences Between Sounds Produced by Different Continuous Interactions. Acoustics 2012, Nantes (2012).
18. Smith, J.O.: website, `https://ccrma.stanford.edu/realsimple/faust_strings/Pick_Position_Comb_Filter.html`
19. Stoelinga, CNJ and Hermes, DJ and Hirschberg, A. and Houtsma, AJM.: Temporal Aspects of Rolling Sounds: A Smooth Ball Approaching the Edge of a Plate. Acta Acustica united with Acustica, 89, 5, 809–817 (2003)

EarGram: an Application for Interactive Exploration of Large Databases of Audio Snippets for Creative Purposes

Gilberto Bernardes[1], Carlos Guedes[1], and Bruce Pennycook[2]

[1] Faculty of Engineering of the University of Porto, Portugal
{g.bernardes, cguedes}@fe.up.pt
[2] University of Texas at Austin, USA
bpennycook@mail.utexas.edu

Abstract. This paper outlines the creative and technical considerations behind earGram, an application built as a Pure Data patch for real-time concatenative sound synthesis. The system encompasses four generative strategies that automatically re-arrange and explore a database of descriptor-analyzed sound snippets (corpus) by rules other than its original temporal order into musically coherent outputs. Of notice are the system's machine-learning capabilities that reveal musical patterns and temporal organizations, as well as several visualization tools that assist the user in making decisions during performance.

Keywords: Concatenative sound synthesis, recombination, and generative music.

1 Introduction

Composing music using audio samples can become a very laborious task. Current solutions that usually involve the use of a music sequencer demand a considerable amount of time to segment and assemble a collection of samples together. During the last decade, a technique called concatenative sound synthesis (CSS) [1] eases the process of synthesizing new sounds based on preexisting audio samples that was extremely difficult and time-consuming when drawn by hand. Concisely, CSS uses a large database of segmented and descriptor-analyzed sound snippets to assemble a target phrase according to a proximity measure in the descriptors space. It was originally intended for text-to-speech synthesis [2], but, later, it was introduced to several other fields that use sound synthesis techniques. CSS is beginning to find its way in musical composition and performance since 2000 [3, 4]. However, the vast majority of literature about this technique still focuses on solving technical problems that enhance the efficiency of these systems, paying very little attention to its musical applications.

The application reported here, i.e. earGram, is a Pure Data (PD) patch that implements a CSS engine and several exploratory tools for musical creative practices. earGram automatically re-arranges sound snippets and permit rapid prototyping of interactive music systems. Particular attention was given to the definition of target phrases to be synthesized, by designing GUIs that allows the user to specify targets quickly and intuitively. Four methods to recombine automatically the units are

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

proposed. They approach two different generative music strategies. The first uses the corpus to synthesize targets defined by an imposed metric and harmonic templates selected beforehand by the user. The second creates a novel music output while retaining the time-varying acoustic morphologies of the audio source(s). Of particular interest is the system's ability to cluster units into representative groups (sub-corpus). The user can control the system in real-time and adjust several structural elements of the output such as the meter, the key, the number of voices, and tempo.

Compared to other CSS software implementations, earGram has three more algorithms related to visualization of the corpus. The common visualization tools that CSS software implementations offer, such as 2d-plots and similarity matrices, depict the distribution and relation amongst units. Eargram offers two more methods that have never been used on the music field and focus on the visualization of high-dimensional data, such as the feature vectors that represent the units, and a third one that aims at depicting the long-term structure of the audio sources(s). They are respectively parallel coordinates [5], star coordinates [6], and arc diagram [7].

Our approach to CSS is inspired on T. Jehan's Skeleton [8] and D. Schwarz's cataRT [9]. The architecture and the conceptual approach of the two systems was our fundamental basis. The analysis-synthesis models presented by Jehan [8] and implemented in Skeleton, especially the perceptual and structural modeling of the music surface, was of seminal importance for the development of the machine listening and learning in earGram. Schwarz's cataRT was equally important due to the similarities of the programing environment used, and its real-time capabilities. earGram extends previous research on generative strategies that recombine descriptor-analyzed units into coherent musical outputs suitable for both studio and live experimentations. In addition, the system offers visual representations of the corpus that were never used in CSS software.

2 System Design

In this section we provide an overview of the design scheme of earGram that is also shown in figure 1.

The user must first feed the system with audio data, either from a live audio input or from pre-recorded material. The first and left-most block in figure 1 (analysis) is responsible for 3 tasks: (1) segmenting the audio material, (2) reducing the content of each unit to a feature vector, and (3) model the harmonic, timbre and metrical structures of the audio source(s) over time. At this stage, a list of pointers to audio segments and their respective feature vector are stored in a database. Subsequently, the system groups the units using one of the available clustering algorithm, and displays the result on a 2d-plot on the main interface of the system (see figure 2).

After importing the audio and analyzing it, the user must choose a generative method for the performance. The available generative methods are responsible for defining a target phrase and retrieving the units that best match this target. At runtime, a signal-processing block enhances the concatenation and emphasizes the artistic expression (right-most bock on figure 1). Among the available audio processing techniques are adaptive filtering, reverberation, chorus, and spectral shift (see sections 7 and 8 for a detailed description).

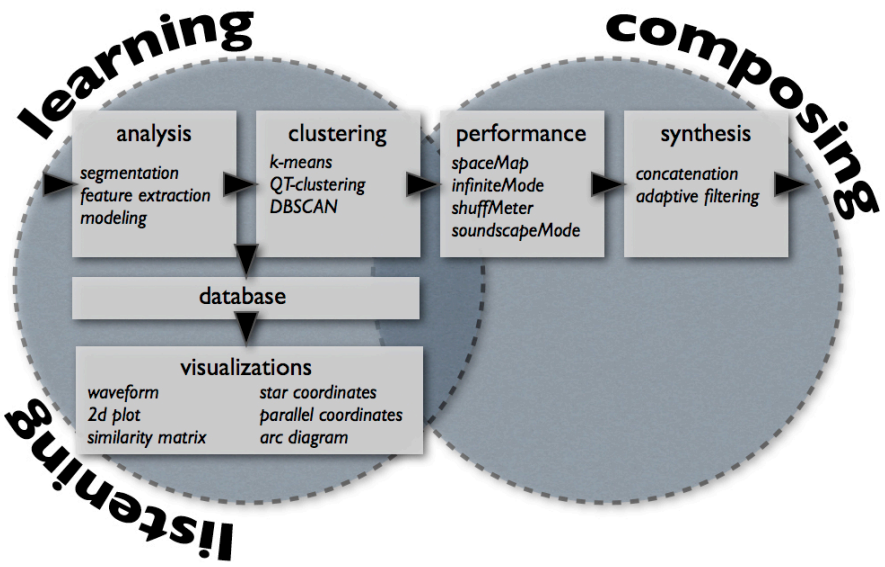


Fig. 1. Design scheme of earGram.

3 Initialization

Initially the user must either create a new project or open a previously saved one, and specify the audio source(s) that will feed the system. There are three options available: (1) single audio track; (2) multiple audio files in a folder; or (3) a live input signal.

The functionality and interface of the proposed system was designed so that musicians that aren't familiarized with MIR research or technology can easily generate some consistent musical results. By default, the system assumes an automatic configuration that needs little to no fine-tuning. However, most settings can be configured via the preferences panel reachable through the main interface. In the following sections, we will describe the system in detail pointing out the differences between the auto-assigned and user-defined settings.

4 Analysis

The analysis block is responsible for three tasks: (1) **segmenting** the audio into units according to a predefined method, (2) defining a **feature vector** that characterizes

each unit, and (3) and **modeling** the units' structure over time regarding harmony, timbre and meter.

In order to **segment** the audio samples using the default settings, the system will inspect the audio input for peaks with harmonic relationships on the spectral flux auto-correlation function to define a regular pulse and segment the audio accordingly. If no such peaks are found, the system will segment the source(s) on every detected *onset*. The user can also assign the segmentation mode manually overwriting the default configuration. Besides the *beat*, and *onset* segmentation methods, there are more methods available: the *uniform* method that segments the audio uniformly based on the length of the window size specified in the system preferences; and the *pitch* method that segments the audio based on the presence of different fundamental frequencies (to be used on monophonic sound sources only). The beat-tracker algorithm used is largely based on S. Dixon [10], and the onset segmentation algorithm is based on P. Brossier [11].

The second purpose of the analysis block is to create for each unit a **feature vector** that represents it. We rely on the *timbreID* library developed by W. Brent [12] for PD to describe the low-level spectral qualities of each unit. We chose this library for its robustness, efficiency, and ability to work in both real time and non-real time. It implements a vast collection of low-level spectral audio descriptors available, such as bark, bfcc, cepstrum, centroid, kurtosis, flatness, flux, irregularity, mfcc, rolloff, skewness, spread, and zero-crossing rate. Two additional low-level features, loudness and fundamental frequency, are extracted by *sigmund~* a PD built-in object created by M. Puckette. Additionally, we built some PD abstractions that define some mid-level features of the input signal, such as tempo, meter, harmonic progressions, and key. Based on this set of descriptors, we represent the units in two ways: (1) static, i.e. constant over the length of the units or (2) dynamic, i.e. varying over the length of the unit.

In the third part of the analysis, the system also creates statistical **models** that represent the harmonic and timbre temporal evolution of the audio source(s). For both music characteristics (harmony and timbre) a transition probability table is created that represents the probability of going from unit *i* to unit *i+1*. The set of all states and transition probabilities completely characterizes a Markov chain, which later allows the generation of new sequences based on stochastic processes. In order to create a transition probability table for harmony and timbre we needed to classify each unit into a finite number of predefined classes. The unit's harmonic content is characterized by the pitch class profile (0-11) of the fundamental bass. The timbre is characterized by a single integer that represents the 3 highest bark spectrum bins, out of a total of 24 bins. Initially, the 3 highest bins are ordered from the lowest to the highest and converted into binary representation. Then the second and the third bins numbers are shifted left by 5 and 10 cases respectively. The three numbers are re-converted to decimal and summed.

Finally, if the input signal was segmented on a beat basis, we build a template that represents the distribution of the units' noisiness for the length of a measure. Zero-crossing rate is a good indicator of the signal noisiness. Very high values denote a very noisy signal while speech or music signals tend to have a very low value. Given the estimated meter with *n* beats per measure, we built a template with *n* bins that represents the noisiness of each beat within a measure. We fill the template by finding the mean value of the zero-crossing rate of all units labeled with a particular pulse,

and repeat the operation for all pulses within the measure. At last, the template is normalized to the range 0-1.

4 Database

A database is created to store the data produced during analysis. It is implemented in PD as a collection of arrays. Each individual array stores the data correspondent to a particular feature for all units in the corpus.

The database and the variables used for the analysis of the sources(s) can be saved as a text file and loaded later, in order to not repeat the time consuming tasks of the database construction, especially if we are dealing with hundreds or thousands of units.

5 Clustering

Clustering aims at grouping similar segments together to form collections of units whose centroid or representative characterizes the group, revealing musical patterns and a certain organization of sounds in time that can be applied in various manners during performance. The current implementation comprises three non-hierarchical clustering algorithms: *k*-means, quality-threshold clustering (QT-clustering), and DBSCAN. We chose this set of algorithms because we considered that they form a good collection to explore the database. If the user wants to have a concise number of clusters defined a priori and consider all units in the corpus, in order to create sub-corpus for different layers or sections, the choice should fall on *k*-means. On the other hand, if the user wants to define the quality of the clusters based on threshold of similarity or neighborhood proximity between units, he/she should choose either QT-clustering or DBSCAN, respectively. The distance metric used to calculate the similarity amongst units in all clustering methods is the Euclidian distance. Even if the clustering algorithms implemented in earGram can deal with arbitrary long vectors, to convey a clearer and more understandable visualization in two dimensions, the algorithms process only two-dimensional vectors selected from the available bag of descriptors.

K-means is one of most popular clustering algorithms available. It partitions the corpus into clusters by allocating each unit to the cluster with the nearest centroid. The total number of clusters *k* needs to be defined a priori. However, the *k*-means implementation in earGram suggests to the user the optimum number of clusters using a technique known as ‘elbow method’. Our implementation of the technique follows two steps. First, we calculate the distortion, i.e. sum of the squared distances between each unit and its allocated centroid for each different value of *k*, ranging from 2 to 9 clusters. Second, we assign the parameter *k* to the number of clusters that doesn't give much better modeling of the data according to a threshold.

QT-clustering was developed by L. Heyer, S. Kruglyak, and S. Yooseph [13] to cluster gene expression patterns. Quality is defined by the cluster diameter and the minimum number of units contained in each cluster. The two parameters are assigned

initially by the user. However, the user does not need to define the number of clusters. All possible clusters are considered: a candidate cluster is generated with respect to every unit and tested in order of size against the quality criteria. In addition, it points out the outliers that should be treated differently (notably excluded) at runtime.

DBSCAN defines the clusters based on the neighborhood proximity and the density of the units in a cluster. Our implementation follows the algorithm described in [14] by M. Ester, H. Kriegel, J. Sander, and X. Xu. The user must define initially two parameters. They are respectively the neighborhood proximity threshold and the minimum density within the radius of each unit. Similarly to the QT-clustering algorithm, DBSCAN avoids defining a priori the number of clusters. However, the algorithm finds arbitrarily shaped clusters very diverse from the ones found by the QT-clustering. It can even find clusters surrounded by (but not connected to) a different cluster.

6 Visualization

Given the huge amount of information that the software produces during analysis, we appended a visualization block to communicate clearly and effectively the information concerning the audio source(s). It is aimed to assist the decision-making during the performance. Most of the visualization tools are interactive and besides displaying the information to the user, they assist him in defining regions of the source material that they want to work with based on structural similarities. The implemented tools and algorithms for data visualization focus on different musical hierarchical levels that demonstrate structural properties of different types. We can roughly divide them in four categories organized from the lowest level, which consists of continuously variable expressive properties to the top levels, which encompass discrete canonical properties: (1) the waveform display is one of most common visualizations tools for audio data, and it provides the user with a general overview of the source(s)' content, and its segmentation; (2) the similarity matrix and the arc diagram [7] aims at presenting the long-term structure of the corpus; (3) the 2d-plots and star coordinates [6] reveal a concise representation of the units based on various combinations of descriptors; and (4) parallel coordinates [5] examines the high-dimension descriptors space.

The waveform plot helps the user to identify and browse through the segmented units.

The self-similarity matrix and arc diagram displays give the user a better understanding of the long-term structure of the audio data by finding similar patterns along the source(s). They depict pairwise similarity between the units of the corpus. The user can group and select different sections on each representation that are treated as different layers during the performance. Non-uniform units require special attention when being compared, because most audio features aren't insensitive to the length of the unit. Thus, a high variability in the unit's length can lead to misleading results on both self-similarity matrix and arc diagram.

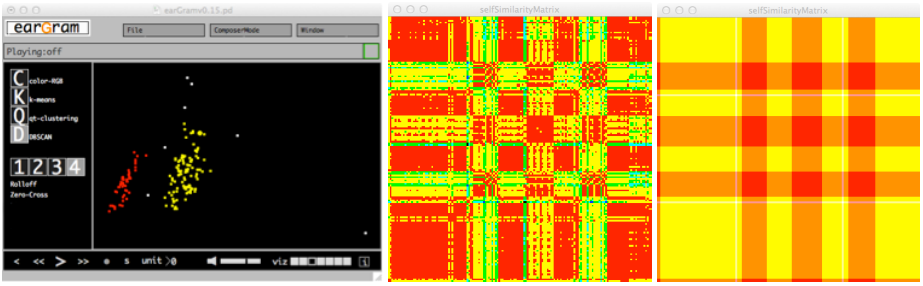


Fig. 2. Different visualizations of a single-track source corpus – *4* by Aphex Twix. From left to right: earGram interface depicting the corpus clustered by a DBSCAN algorithm on a 2d-plot (using spectral rolloff and zero-crossing rate as variables). Middle and right images are self-similarity matrices plotting the same corpus. The middle image depicts the similarity using all available descriptors, and the right most uses the color scheme gathered from the cluster representation on the interface.

The 2d-plot is one of most common visualizations adopted by CSS software. It is especially suitable for navigating and exploring the corpus intuitively. Similar units are plotted together, and its representation along the axis reveals characteristics related to the axis' variable. Additionally, another layer of information concerning the units' color is also available. The color of each unit is defined by a list with three elements that correspond to the red, the green, and the blue values of an additive (RGB) color model. The values that compound the list that defines the units' color are chosen from the available audio descriptors. Star coordinates is a dimensionality reduction algorithm presented by E. Kandogan [6]. It maps high-dimensional data linearly to 2d or 3d using their vector sum. Here we used this algorithm to represent multiple features on a 2d representation. We chose this algorithm for its understandability (each dimension still preserves the same meaning), contrary to approaches such as multidimensional scaling or principal component analysis. One disadvantage of star coordinates is the need to explore the representation by weighing the variables and assigning the axis to different angles to find interesting patterns. Parallel coordinates [5] is barely used in the music domain but is a known procedure to visualize high-dimensional data and analyze multivariate data. By default, all descriptors are taken in account to create the projection, although the user can select the features he/she wants to include.

Figure 2 depicts three representations of the same corpus that comprises a singe audio track – *4* by Aphex Twin. The structure of the song is clarified by the matrix in the rightmost image, which represents the units by the colors resulting from the DBSCAN clusters (leftmost image). We can clearly notice that the song comprises two sections that alternate throughout the track.

7 Performance

The main drive behind the analysis is primarily synthesis. On the following sections, we present four methods that re-arrange in a structured and musical meaningful way

the collection of units that form the corpus based on the analysis described in the previous sections. The recombination processes cover the generation of three specific music results: (1) sonic textures / soundscapes (space-Map and soundscapeMode) either by browsing the navigable 2d visualizations or by defining targets according to audio qualities, (2) extending indeterminately the length of a particular audio sample avoiding repetitions (infiniteMode), and (3) defining targets that reflect a particular meter (shuffMeter).

The methods described bellow are both responsible for defining the target phrases and selecting the units that best matches the target queries.

7.1 SpaceMap

This method is meant to function as a tool that allows intuitive and interactive exploration of the units on the 2d-visual representation. It can be seen as an extended granular synthesis engine where grains are organized in a meaningful visual representation. It aims at creating sonic textures with controllable nuances. It is a very powerful method when playing along with a live input source particularly when improvising, because besides the automatic and meaningful segmentation that the software produces, after a segment is defined it is consequently plotted in the interface, creating an almost instantaneous representation of the input signal during performance.

It has three playing modes: (1) mouseOver – continuously maps the mouse position on screen to the granulator's parameters; (2) pointerClick – the same effect as mode 1, but only when the mouse button is pressed a unit is played; and (3) colorPicker – selects units based on their RGB color values that are retrieved from a navigable grid of colors.

Several parameters can be changed during performance and affect each unit separately, such as gain, density of events, pitch deviations, and panning. All parameters can have a certain degree of random variability. The software also allows the creation of several bus-channels that may incorporate audio effects. At runtime, the representation of the units in the interface can be changed without affecting the synthesis, except when a live input source is fed to the system.

7.2 InfiniteMode

The second synthesis method implemented in earGram aims at generating a musical result of indeterminate length that doesn't repeat, while it retains the time-varying acoustic morphologies of the audio source(s). It is primarily suitable for single-track audio input, or for groups of sound files that share commonalities at the metrical, harmonic and timbre level.

Each new unit is triggered and defined at the end of the previous one and defined at that stage based on the harmonic, metrical, timbre and noisiness models created beforehand during analysis. The interface allows us to select and use up to three sets of characteristics that will be responsible for defining the target. On the interface, two sets of characteristics are predefined: one for soundscapes (timbre) and a second for polyphonic music (meter, harmony and timbre).

At runtime, the algorithm selects a new unit to synthesize by finding all units that both satisfy the assigned group of characteristics and that best matches the spectrum of the previous unit, i.e. when a new unit is triggered, the algorithm examines all selected characteristics, and for each of them defines a group of units that match the query. Then, it finds the units that are common to all groups of characteristics, and finally, from the remaining units is selected the one that minimizes the distance on the bark spectrum representation to the previous selected unit. If the algorithm doesn't find any unit that satisfies all the assigned characteristics the algorithm will ignore sequentially characteristics until finding candidates by the contrary order of the interface. If we have three selected characteristics, and any unit is found for a specific query, the algorithm eliminates the third characteristic and examines again the number of remaining units, if nothing is retrieved it eliminates then the second and so on.

Harmony and timbre aims at preserving the temporal evolution of chord progressions and audio spectra from the original source(s). At every new query a group of units is outputted for each element, according to the transition probability table elaborated during analysis.

To preserve the metrical accents' distribution over the length of a bar during synthesis, the algorithm retrieves for each metrical accent the units that were previously labeled accordingly during analysis. In other words, initially, while at the original temporal units order, every unit is labeled with their respective position over the length of a bar, in a sequence that goes from 1 to number of units per bar. E.g., assuming we got a time signature with four units every bar, we would split the source in groups of four units, and label each sequentially. At runtime, for each new metrical accent, the algorithm retrieves all units that were labeled with that metrical accent.

The noisiness characteristic attempts to replicate the zero-crossing rate configuration over the length of a measure that was encoded in a template elaborated during analysis. At each query, successive values are retrieved from the template and the algorithm looks at the database to find units that present a similar zero-crossing rate value. Given the template value x the algorithm retrieves all units that fall on the interval $[x\text{-}0.1, x\text{+}0.1]$.

7.3 ShuffMeter

Clarence Barlow's metric indispensability principle [15] has been successfully applied as a metrical supervision procedure when generating drum patterns in a particular style [16] as well as a model for constraining a stochastic rhythmic generation algorithm given a particular time signature [17]. The two algorithms work with symbolic music representations. shuffMeter extends previous research by applying Barlow's principle to drive the definition of targets that reflect a particular meter.

Given the scope of this paper and space restrictions, we cannot detail all the implementation of Barlow's metric indispensability. However, we follow the implementation described in [16].

After assigning a meter and a specific metrical level, the algorithm defines a hierarchical organization of the strong and weak beats of the meter to be better perceived by a listener. We mapped the weights into two audio descriptors: loudness

and spectral flux, by assuming that loudness and spectral changes are most likely to occur on the strongest meter accents. To simplify the computation we merged both descriptors into a single descriptor defined as their mean value. For each query the algorithm gathers the metrical weight *w* for that specific accent, and retrieves all units from the corpus with a value of *w*±0.1 for that descriptor.

We can apply this principle either on the whole corpus or on separate clusters, allowing as many layers as the number of existing clusters. The user can navigate in real time in a two dimensional map in the form of a square. Two pairs of variables mapped to each of the vertices of the square will adapt the configuration of the weights. The horizontal direction, from rough to smooth, will regulate the variability between all accents. The vertical direction, from loud to soft, will increase or diminish the weights proportionally.

Each concatenated unit is triggered by a timer assigned to the duration correspondent to the current beats per minute (bpm). This method was adopted here instead of a more natural strategy implemented in the previous section (7.2. infiniteMode), given the need to synchronize several units with slightly different lengths. If the units' length doesn't match the specified duration, they are consequently scaled in time by recurring to a time-stretch algorithm, which changes the speed of the audio signal without affecting the pitch.

7.4 SoundscapeMode

The soundscapeMode is a recombination method that was specially designed to recreate and work with environmental sound sources. It is a valuable and easy tool to design sound for film or installations, since it can structurally arrange on a map the units according to their perceptual qualities. The map has the form of a square divided into four main regions arranged in pairs of interconnected variables. The user can navigate in real time on the map as if he would navigate through a sound cartography. The first variables' pair controls the density of events (dense and sparse) and the second the roughness of the events (smooth and sharp).

The horizontal variable is density, i.e. the number of units played simultaneously, and ranges from 1 to 5. Smooth-sharp dichotomy, the second variables' pair represented vertically, aims at regulating and organizing the corpus in terms of diversity and stability and it is driven by the spectral flux descriptor. Spectral flux is a frequency domain feature, and describes the fluctuations in the spectrum of the signal. It was chosen because it is powerful in denoting attacks and sudden changes in the spectrum and thus for showing how stable the audio is along the unit. It is prudent to note that the application is highly dependent on the source file(s). If we feed the system with varying texture samples, the difference between smooth and sharp will be almost imperceptible.

As in infiniteMode, we added a block at the end of the target's definition that intends to maintain the best possible continuation between concatenated units, in terms of loudness and spectral changes. It is done by gathering all units' candidates for a specific query and finding the one that minimizes the distance on the bark spectrum representation to the previous selected unit.

8 Synthesis

Synthesis is done by concatenating units with a slight overlap. Each unit is played with amplitude envelope in the shape of a bell curve.

Most recombination methods make sure that the best possible continuity between concatenated units is guaranteed – i.e. if more than one unit matches the target at a certain point of the phrase, the system will select the unit that best matches the spectrum of the previous one. However, discontinuities and gaps still occur. To improve the quality of the synthesis an additional feature is added at the end of the chain in order to filter certain transition discontinuities on the audio flow. This is done with the help of an external object from the soundhack plugins bundle [18] named +spectralcompand~, which is a spectral version of the standard expander/compressor, commonly known as compander. It divides the spectrum in 513 bands and processes each of them individually. The algorithm computes iteratively the spectrum every 50 ms and applies it as a mask during synthesis.

9 Applications

The four recombination algorithms detailed in section 7 are suitable for a variety of music situations, spanning from sound installations to concert music. The design of the system doesn't reflect any particular music style. Our main purpose was to design a music system that learn from the music it draws its database from, and define coherent target phrases to be synthesized. Thus, the music output is highly dependent on the sound source(s), which are entirely selected by the user. In addition, some user supervision is needed to select certain recombination methods over others given the nature of the sound source(s). For example, if we fill the database with polyphonic music signals segmented on a beat basis it will be highly implausible that this collection of units will produce a consistent result when using soundscapeMode, which is mainly intended to recombining environmental sounds.

The system is easily adjustable to the context of interactive performance. All recombination methods have some degree of variability that can be easily controlled on the GUI. The interface for all recombination methods is intuitive and built as navigable maps that are almost self explainable. Instead of adjusting manually the several variables, the user can map characteristics extracted from an ongoing performance, whether they are motion, or sonic characteristics, or even any other measurable features extractable from a particular setting to any controllable variable of the interface.

The main purpose of the machine listening and learning techniques implemented in earGram is to drive the synthesis part of the software. However, the system may be useful for other applications domains outside this realm. The analytical and visualization tools that the software provides may constitute a valuable resource for the purpose of analyzing music under several fields such as computational musicology and cognitive musicology.

9 Conclusions and Discussion

This paper presents earGram, a novel CSS application built in Pure Data that comprises four generative music strategies that re-assign the original temporal order of the corpus for interactive music contexts, focusing on the user interface, MIR techniques of data analysis, mining and retrieval, and probabilistic modeling of timbre and harmony.

The visual representations offered in earGram gives the user a better understanding of the entire collection of units and the similarity amongst them. Most visualizations also allow interactive and guided exploration of the corpus, suitable for creating soundscapes and elucidating some decision-making concerning performance. The use of Barlow's indispensability algorithm proved to be an efficient method to ensure metrical coherence during the recombination process by providing a template that guides the definition of targets according to a predefined meter. A Markov chain algorithm was successfully applied to generate an infinite number of variations on the original signal with a minimum amount of interaction, while retaining the time varying morphologies of the source(s) modeled by a transition probability table between units.

The software together with many sound examples for all the recombination methods detailed in the paper and their respective project template used to create the examples are available at: https://sites.google.com/site/eargram/.

10 Future Work

The audio source(s) used during analysis are determinant for the possible outputs the system can consistently offer. CSS is only as good as the database from which it draws its sound units. Thus, besides requiring a rich database of sounds, more research should be addressed towards a better understanding of the source(s), which would contribute for more refined way of using the descriptors and help restricting the application field.

Concerning analysis further additions that center on the rhythmic content of the audio source(s) are under development. We believe that a better understanding of the source's rhythmic structure will enhance solidity during performance, particularly by avoiding gaps in the synthesis continuum, and by favoring typical articulations and textures of the corpus. Still regarding rhythmic instabilities, when layering different clusters of units, as it is done most notably in shuffMeter, but also in soundscapeMode, some rhythmic incongruence arouse due to the lack of alignment between overlapping units. Besides the valuable contribution that the before-mentioned rhythmic descriptions can add, some experiments will envisage to time-stretch units to align their rhythmic content.

Flexible methods for sequencing and mixing different recombination methods, various clusters, and diverse corpuses (notably combining corpus from different sources, e.g. live and pre-recorded sound) are under development.

Finally, concerning the applications domain, we consider that evolutionary methods could help at orienting the system, notably by defining larger targets that

consider more than the transition between consecutive units and by allowing control over the evolving process.

Acknowledgments. This work was partly supported by the European Commission, FP7 (Seventh Framework Programme), ICT-2011.1.5 Networked Media and Search Systems, grant agreement No 287711 (MIReS). We would like to thank George Sioros for his careful review of this paper.

References

1. Schwarz, D.: Current Research in Concatenative Sound Synthesis. In: Proceedings of the International Computer Music Conference, Barcelona, Spain (2005)
2. Hunt, A. J., Black, A. W.: Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (1996)
3. Zils, A., Pachet, F.: Musical mosaicking. In: Proceedings of the COST G-6 Conference on Digital Audio Effects, Limerick, Ireland, December (2001)
4. Schwarz, D.: Musical Applications of real-time corpus-based concatenative synthesis. In: Proceedings of the International Computer Music Conference (2007)
5. Inselberg, A.: Parallel Coordinates: Visual Multidimensional Geometry and Its Applications. Springer (2009)
6. Kandogan, E.: Visualizing Multi-dimensional Clusters, Trends, and Outliers using Star Coordinates. In: Proceedings of the Knowledge and Data Mining (2001)
7. Wattenberg, M. Arc Diagrams: Visualizing Structure in Strings. In: Proceedings of the IEEE Information Visualization Conference (2002)
8. Jehan, T.: Creating Music by Listening. Ph.D. Thesis, M.I.T., MA (2005)
9. Schwarz, D., Cahen, R., Britton, S.: Principles and Applications of Interactive Corpus-based Concatenative Synthesis. In: Journées d'Informatique Musicale, GMEA, Albi, France (2008)
10. Dixon, S.: An interactive beat tracking and visualization system. In: Proceedings International Computer Music Conference (2001)
11. Brossier, P.: Automatic Annotation of Musical Audio for Interactive Applications. Ph.D. thesis, Queen Mary, University of London (2006)
12. Brent, W.: A Timbre Analysis and Classification Toolkit for Pure Data. In: Proceedings of the International Computer Music Conference, New York, EUA (2010)
13. Heyer, L., Kruglyak S., Yooseph, S.: Exploring Expression Data: Identification and Analysis of Coexpressed Genes. Genome Research, 9:1106-1115 (1999)
14. Ester, M., Kriegel H., Sander, J., Xu, X.: A density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: Proceedings of the Knowledge Discovery and Data Mining. AAAI Press, pp. 226–231 (1996)
15. Barlow, C.: Two essays on theory. Computer Music Journal, 11, pp. 44-60 (1987)
16. Bernardes, G., Guedes, C., Pennycook, B.: Style Emulation of Drum Patterns by Means of Evolutionary Methods and Statistical Analysis. In: Proceedings of the Sound and Music Computing Conference, Barcelona, Spain (2010
17. Sioros, G., Guedes, C.: Automatic Rhythmic Performance in Max/MSP: the kin.rhythmicator. In: Proceedings of the International Conference on New Interfaces for Musical Expression, Oslo, Norway (2011)
18. SoundHack Plugins Bundle, http://soundhack.henfast.com/

From Shape to Sound: sonification of two dimensional curves by reenaction of biological movements

Etienne Thoret[1], Mitsuko Aramaki[1], Richard Kronland-Martinet[1], Jean-Luc Velay[2], and Sølvi Ystad[1⋆]

[1] Laboratoire de Mécanique et d'Acoustique
last-name@lma.cnrs-mrs.fr
[2] Laboratoire de Neurosciences Cognitives
last-name@univ-amu.fr

Abstract. In this study, we propose a method to synthesize sonic metaphors of two dimensional curves based on the mental representation of friction sound produced by the interaction between the pencil and the paper when somebody is drawing or writing. The relevance of this approach is firstly presented. Secondly, synthesized friction sounds that enable the investigation of the relevance of kinematics in the perception of a gesture underlying a sound are described. In the third part, a biological law linking the curvature of a shape to the velocity of the gesture which has drawn the shape is calibrated from the auditory point of view. This law enables generation of synthesized friction sounds coherent with human gestures.

Keywords: Sonification - Gesture - Drawings - 2/3-power law - Scraping / Friction sounds - Sound Perception

1 Introduction

The possibility to convey information with sounds has been largely investigated the last thirty years and is now commonly called sonification. This field of research aims at transmitting information by sounds either instead of or in addition to a visual display. A common example is the Geiger-Müller counter which produces clicks depending on the quantity of ionizing radiation. The temporal aspect of sounds is particularly interesting to convey dynamic information which could not have been displayed on a screen or with less accuracy.

Pioneering ideas within the domain of sonification were developed by Gaver who adapted Gibson's ecological theory of visual perception to auditory perception [5] to create sounds from perceptual invariants providing relevant informations about an action. Since then, many studies dealing with applications within

⋆ This work is supported by the French National Research Agency (ANR) under the *MetaSon - Sonic Metaphors* Project – CONTINT 2010 : ANR-10-CORD-010.

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

a large number of fields, such as sport training, industrial processes, medicine have been proposed to convey useful information with sound.

This study is included in a larger research project which explores the possibilities to create sound metaphors in the context of different applications[3]. One of these applications is the rehabilitation of dysgraphic children[4] with the use of sounds to guide them to recover the right handwriting gesture. To achieve this goal, we first need to understand how a gesture could be perceptually linked to a sound, and which sound attributes can be used to inform of the dynamical characteristics of the gestures.

In this article, we aimed at proposing a synthesis tool to sonify drawings and more generally two-dimensional shapes. We hereby considered a sonification strategy based on the evocation of the underlying human gestures that might have produced the shapes. In other terms, we aimed at sonifying a drawing by virtually re-enacting a natural gesture of a human that drawn the shape. We considered sounds naturally generated by the interaction between a pencil and a rough surface during the drawing process, i.e. friction sounds. To support our approach, we investigated the relationship between a sound and the evoked gesture and whether a sound can inform of the drawn shape. We therefore conducted experiments to highlight the relevance of the velocity profile as a perceptual attribute of sound that convey information on the underlying gesture and on the drawn shape.

The article is organized as follows. The relationship between a drawn shape and the generated friction sound is firstly studied. We designed a listening test based on a shape/sound association task aiming at examining the subjects ability to recover the correct drawn shape from the sound only. Then, we investigated the influence of the velocity profile on the perceived gesture and shape. For that, a simple synthesis model of friction sounds was used to control this parameter independently from the other ones that are present in a natural gesture (such as velocity, pressure, pencil orientation...). In a third part, the possibility to re-generate a *human* velocity profile of a gesture from the geometrical characteristics of a shape is investigated by a listening test that consisted in calibrating a biological law linking the curvature of a shape to the kinematics of a gesture. Based on this results, a sonification process of shapes is proposed.

2 Shape Discrimination from Friction Sounds

To our knowledge, the relationship between the sound and the drawn shape was not formally investigated in the literature from a perceptual point of view. We therefore designed an experimental protocol aiming at better understand this relationship.

When somebody is drawing, the sounds produced by the friction between the pencil lead and the paper are linked to the gesture behind the drawing. In this

[3] `http://metason.cnrs-mrs.fr/`

[4] Dysgraphy is a motor problem which consequences are difficulties with graphic gestures and to write.

study we examine whether these sounds convey information about the shape which is being drawn.

Stimuli were obtained from recordings of friction sounds produced during a drawing process. A person was asked to draw six predefined shapes (Circle, Ellipse, Loops, Lemniscate, Line, Arches) on a graphic tablet. The velocity profiles of the writer's gestures were also recorded during this process.

To evaluate the possibility to reveal a shape from the friction sounds, a listening test was then set up where subjects were asked to associate one of the recorded sounds to one of the drawn shapes [7]. From the six shapes recorded on the writer, two corpuses of four shapes (two shapes were common between corpuses) were defined; one with very distinct shapes and one with more similar shape, see Figure 1. For each corpus, the subjects were asked to univocally associate one friction sound (among the four available) to one shape.

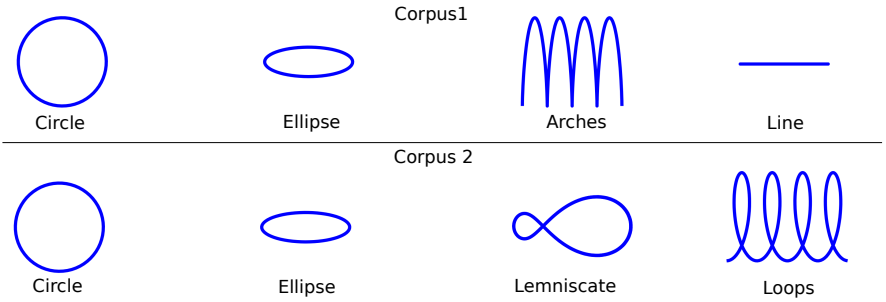


Fig. 1. The two corpuses of four shapes of the association tests

The results of the test show that, except for the *Loops*, each sound was associated with the correct shape with a success rate above random level[5].

In the case of the first corpus, every sounds were properly associated to the shape. The rates of success were: Circle: 98.75% – Ellipse: 81.25% – Arches: 80% – Line: 87.5%.

In the case of the second corpus, confusions appear between the *Ellipse* and the *Loops*, and only the *Loops* were not recognized above chance. The rates of success were: Circle: 97.22% – Ellipse: 41.67% – Lemniscate: 68.06% – Loops: 29.17%.

Although some confusions occurred between shapes of the second corpus, we obtained relatively high success rates. These data showed that sounds produced during the drawing contain accurate information about the drawn shape. To determine the acoustical characteristics that convey this information, we further investigated the relevance of the velocity profile that is one of the important parameter of the motions dynamics.

[5] The random level is defined at 25% sound to shape association rate.

3 Perceptual Relevance of the Velocity Profile

To focus on the influence of the velocity profile, we used a synthesis model which gives the possibility to synthesize friction sounds from the velocity profiles previously recorded on the writer (section 2) by fixing the other parameters (such as pressure, pencil orientation) as constant. We also assumed that the nature of the rubbed surface was identical. A same shape/sound association test as the previous one was conducted with synthetic friction sounds to investigate the perceptual information provided by the velocity profile only. In the following sections, the synthesis model of friction sound is firstly presented and then results of the listening test are discussed.

3.1 A physically based model of friction sounds

Friction sounds have been largely studied and have been the subject of a wide number of applications in different domains of physics. Here we present a simple and common model of friction sounds based on a phenomenological approach of the physical source. This model was firstly presented by Gaver in [4] and improved by Van den Doel in [8].

When a pencil is rubbing a rough surface, the produced sound could be modeled as successive impacts of the pencil lead on the asperities of the surface. With a source-resonator model, it is possible to create friction sounds by reading a noise wavetable with a velocity linked to the velocity of the gesture and filtered by a resonant filter bank adjusted to model the characteristics of the object which is rubbed, see Figure 2. The noise wavetable represents the profile of the surface which is rubbed. Resonant filter banks simulate the resonances of the rubbed object and are characterized by a set of frequency and bandwidth values. Previous studies proposed some mapping strategies allowing for a control of these synthesis parameters based on perceptual attributes (such as the perceived material or size) [1, 2].

3.2 Test and Results

The previous synthesis model allowed us to generate synthetic sounds from a given velocity profile and to accurately investigate whether this parameter is a relevant characteristic of sound perception. We used the velocity profiles previously collected on the graphic tablet and we designed a mapping between these profiles and the cutoff frequency of the lowpass filter. The same listening test as the one presented in section 2 was carried out. Results showed that the shapes of the first corpus (distinct shapes) were properly associated with the sounds with high success rates. The shapes of the second corpus (similar shapes) were associated with lower success rates than for the first corpus, but always above chance level.

In addition, results showed a lack of significant differences between the two experiments (analysis conducted with the type of sounds as factor: recorded vs synthetic sounds). These results revealed that sounds computed from the

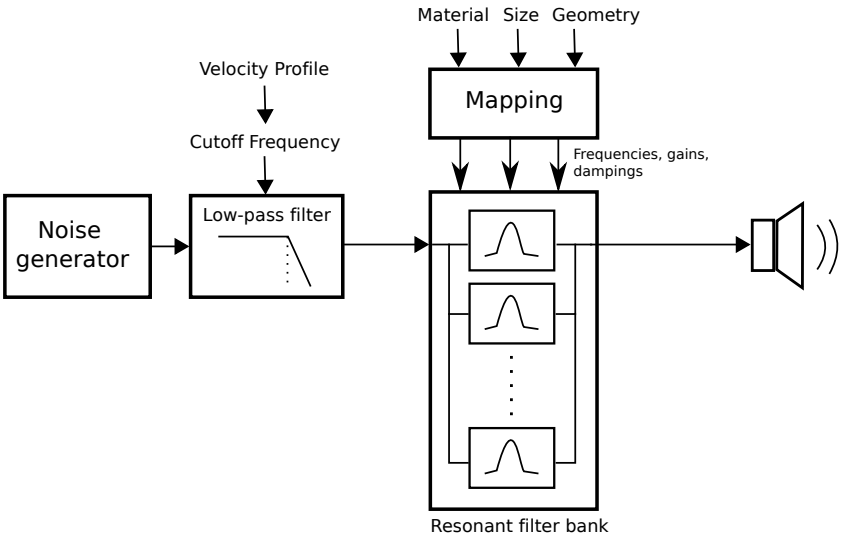


Fig. 2. Physically based friction model

velocity profiles provided as much useful information for shape recognition as the recorded ones. This means that the velocity profile contains the information needed on a shape.

4 Sonification Strategy of Human Drawing

The previous sections highlighted that a mental representation of a shape can be elicited from the sound produced when this shape is drawn and that the velocity profile is a relevant feature of the gesture to convey information on this drawn shape.

In this section we propose a sonification strategy of a drawn trace by recovering the human gesture that produced the trace. We want to create a sound from a given shape using the previous friction sound synthesis model, and the velocity profile as a control parameter. For that purpose, the velocity profile is estimated with respect to the geometrical characteristics of the shape.

4.1 A biological law of motion for the drawing gestures: the 2/3-power law

To regenerate a velocity profile from a given shape, we referred to a biological law which linked the radius of curvature R_c of a shape to the tangential velocity v_t of the gesture which drew it. In [6], Viviani highlighted this relation called the 2/3-power law which expressed the covariations of these two variables with the following formula:

$$v_t(s) = KR_c(s)^{1-\beta} \tag{1}$$

with $\beta = 2/3$, K is assumed to be constant.

The relevance of this law with respect to the motor competences such as drawing and more generally in many natural movements has been largely studied [6, 10].

This law has also been highlighted in perceptual processes. In the case of visual perception, a study revealed that the perception of the velocity of a point moving along a curved shape should be modulated by such a power law so that the velocity of the point is perceived as constant when the exponent is equal to 2/3 [9]. It means that the notion of perceived constant velocity is not a associated to a physical constant velocity, but to a velocity which respect a specific biological constraint, the 2/3-power law.

4.2 Calibration of the 2/3-power law in the auditory modality

In [7], the relevance of this law was investigated from the auditory perception point of view by a calibration test of the exponent β of the equation 1. For that, we used the previous synthesis model of friction sound. The velocity profile was computed by using the 2/3-power law with a fixed mean velocity K, and with a curvature profile which corresponds to a pseudo-random shape (cf. Figure 3) to avoid preferences on specific known shapes. Each subject did 6 trials and a pseudo-random shape was generated at each one. Subjects listened to the corresponding friction sound and were asked to modify the sound (by acting on the β value) until they could imagine that a human has produced this sound by drawing. The initial value of β was randomized at each trial and the shape was not shown to subjects so that they could focus on the sound only.

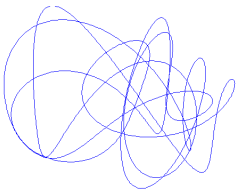


Fig. 3. Example of pseudo-random shape.

We found that the mean value of the exponent was $\beta = 0.64$ ($SD = 0.08$), which means that the most realistic velocity profile which characterizing a human gesture from an auditory point of view follows the 2/3-power law.

This results allowed us to validate the use of the 2/3-power law to generate a velocity profile from a given shape. The obtained velocity profile can further be used to synthesize a sound underlying a mental representation of the gesture.

5 Sonification Tool

The three previous sections gave perceptual results and technical expertise to create a sonification tool of two dimensional curves based on the auditory perception of friction sounds produced by human gestures.

This tool aims at giving a mean to create a sound perceptually coherent with a given shape[6]. The input of this tool could be a scanned shape as well as a shape recorded with a graphic tablet. The Figure 4 sums up the sonification process.

1. The user has to choose a start point on the shape and the direction of the movement
2. From the input shape, the curvature is computed from the coordinates $(x(s), y(s))$ of each point of the shape
3. A velocity profile is created from the curvature with the 2/3-power law
4. The mean velocity of the gesture can be controlled with the coefficient K of the 2/3-power law. The velocity profile controls a friction sound synthesis model and generates a sound coherent with the given shape. The sound could also be played coherently with a displayed movie where the shape is synchronously drawn with the friction sound

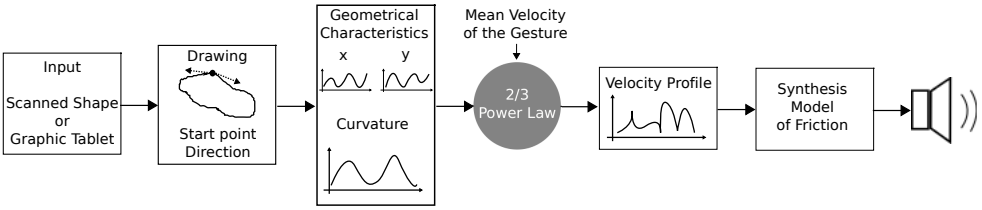


Fig. 4. Complete sonification process

6 Conclusions and perspectives

In this article we proposed a sonification strategy of shapes that could be applied to any set of two dimensional data which could be expressed as a couple of continuous functions. This sonification process, based on the mental representation of a biological gesture underlying a friction sound, transforms the curvature of a shape into a velocity profile which is further used to synthesize realistic friction sounds evoking a gesture coherent with the drawn shape.

This preliminary study also brought up many perspectives. First concerning the possibility to apply the obtained velocity profile to new sound textures other

[6] An example of sonification is available on the following website : `http://www.lma.cnrs-mrs.fr/~kronland/ShapeSoundCmmr`

than friction noise. For instance, if we modulate the pitch of a sound by the velocity profile of a gesture, will this transformation also be relevant for sonifying a shape? More generally, can we use this transformation to create sonic metaphors of a human gesture or drawn shape with abstract sound textures such as wind for example?

Another perspective triggered by this study is the possibility to use the sonification process proposed here for a visual display of a moving spot-light to investigate the multimodal integration of auditory and visual information in the perception of movement dynamics. Viviani highlighted that the 2/3-power law defined a perceived constant velocity in the visual domain. In the auditory domain, we clearly pay attention to *variations* in the sound. It would therefore be interesting to study whether the visual illusion of constant velocity is present when a sound is presented together with the visual display that follows the 2/3-power law.

It should be noted that this work could also be applied to the development interfaces to assist visually impaired. It indeed gives a new way to evoke shapes with sounds.

References

1. M. Aramaki, M. Besson, R. Kronland-Martinet, and S. Ystad. Controlling the Perceived Material in an Impact Sound Synthesizer. IEEE Transaction On Speech and Audio Processing, 19(2) :301–314, February 2011.
2. M. Aramaki, C. Gondre, R. Kronland-Martinet, T. Voinier and S. Ystad. Thinking the sounds: an intuitive control of an impact sound synthesizer. Proceedings of ICAD 09 – 15th International Conference on Auditory Display, 2009.
3. W. W. Gaver. The SonicFinder: An interface that uses auditory icons. Human-Computer Interaction, 1(4) :67–94, Taylor & Francis, 1989.
4. W. W. Gaver. Synthesizing auditory icons, Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems, 228–235, ACM, 1993.
5. J. J. Gibson. The Senses Considered as Perceptual Systems . Boston: Houghton-Mifflin, 1966.
6. F. Lacquaniti, C. A. Terzuolo, and P. Viviani. The law relating kinematic and figural aspects of drawing movements, Acta Psychologica, 54, 115–130, 1983.
7. E. Thoret, M. Aramaki, R. Kronland-Martinet, J. L. Velay, and S. Ystad. Sonification of Drawings by Virtually Reenacting Biological Movements. Versatile Sound Models for Interaction in Audio-Graphic Virtual Environments, Workshop @ DAFx-11, `http://metason.cnrs-mrs.fr/Documents/SonificationOfDrawings/SonificationOfDrawingDafx11.html`, September 2011.
8. K. Van Den Doel, P. G. Kry and D.K. Pai. FoleyAutomatic : physically-based sound effects for interactive simulation and animation. Proceedings of the 28th annual conference on Computer graphics and interactive techniques, 537–544, ACM, 2001.
9. P. Viviani, and N. Stucchi. Biological movements look uniform: Evidence of motor-perceptual interactions. Journal of Experimental Psychology: Human Perception and Performance, 18, 603–623, 1992.
10. P. Viviani, and T. Flash. Minimum-jerk, two-thirds power law and isochrony: Converging approaches to movement planning. Journal of Experimental Psychology: Human Perception and Performance, 21, 32–53, 1995.

Oral session 3:

Computer Models of Music Perception and Cognition: Applications and Implications for MIR

The Role of Time in Music Emotion Recognition

Marcelo Caetano[1]* and Frans Wiering[2]

[1] Institute of Computer Science, Foundation for Research and Technology - Hellas FORTH-ICS, Heraklion, Crete, Greece
[2] Department of Information and Computing Sciences, Utrecht University, Utrecht, Netherlands
caetano@ics.forth.gr, f.wiering@uu.nl

Abstract. Music is widely perceived as expressive of emotion. Music plays such a fundamental role in society economically, culturally and in people's personal lives that the emotional impact of music on people is extremely relevant. Research on automatic recognition of emotion in music usually approaches the problem from a classification perspective, comparing "emotional labels" calculated from different representations of music with those of human annotators. Most music emotion recognition systems are just adapted genre classifiers, so the performance of music emotion recognition using this limited approach has held steady for the last few years because of several shortcomings. In this article, we discuss the importance of time, usually neglected in automatic recognition of emotion in music, and present ideas to exploit temporal information from the music and the listener's emotional ratings. We argue that only by incorporating time can we advance the present stagnant approach to music emotion recognition.

Keywords: Music, Time, Emotions, Mood, Automatic Mood Classification, Music Emotion Recognition

1 Introduction

The emotional impact of music on people and the association of music with particular emotions or 'moods' have been used in certain contexts to convey meaning, such as in movies, musicals, advertising, games, music recommendation systems, and even music therapy, music education, and music composition, among others. Empirical research on emotional expression started about one hundred years ago, mainly from a music psychology perspective [1], and has successively increased in scope up to today's computational models. Research on music and emotions usually investigates listeners' response to music by associating certain emotions to particular pieces, genres, styles, performances, etc. An emerging field is the automatic recognition of emotions (or 'mood') in music, also called music emotion recognition (MER) [7]. A typical approach to MER categorizes emotions into a number of classes and applies machine learning techniques

* This work is funded by the Marie Curie IAPP "AVID MODE" grant within the European Commissions FP7.

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

to train a classifier and compare the results against human annotations [7, 22, 10]. The 'automatic mood classification' task in MIREX epitomizes the machine learning approach to MER, presenting systems whose performance range from 22 to 65 percent [3]. Researchers are currently investigating [4, 7] how to improve the performance of MER systems. Interestingly, the role of time in the automatic recognition of emotions in music is seldom discussed in MER research.

Musical experience is inherently tied to time. Studies [8, 11, 5, 18] suggest that the temporal evolution of the musical features is intrinsically linked to listeners' emotional response to music, that is, emotions expressed or aroused by music. Among the cognitive processes involved in listening to music, memory and expectations play a major role. In this article, we argue that time lies at the core of the complex link between music and emotions, and should be brought to the foreground of MER systems.

The next section presents a brief review of the classic machine learning approach to MER. Then, we discuss an important drawback of this approach, the lack of temporal information. We present the traditional representation of musical features and the model of emotions to motivate the incorporation of temporal information in the next section. Next we discuss the reationship between the temporal evolution of musical features and emotional changes. Finally, we present the conclusions and discuss future perspectives.

2 The Traditional Classification Approach

Traditionally, research into computational systems that automatically estimate the listener's emotional response to music approaches the problem from a classification standpoint, assigning "emotional labels" to pieces (or tracks) and then comparing the result against human annotations [7, 22, 10, 3]. In this case, the classifier is a system that performs a mapping from a feature space to a set of classes. When applied in MER, the features can be extracted from different representations of music, such as the audio, lyrics, the score, among others [7], and the classes are clusters of emotional labels such as "depressive" or "happy". There are several automatic classification algorithms that can be used, commonly said to belong to the machine learning paradigm of computational intelligence.

2.1 Where Does the Traditional Approach Fail?

Independently of the specific algorithm used, the investigator that chooses this approach must decide how to represent the two spaces, the musical features and the emotions. On the one hand, we should choose musical features that capture information about the expression of emotions. Some features such as tempo and loudness have been shown to bear a close relationship with the perception of emotions in music [19]. On the other hand, the model of emotion should reflect listeners' emotional response because emotions are very subjective and may change according to musical genre, cultural background, musical training and exposure, mood, physiological state, personal disposition and taste

[1]. We argue that the current approach misrepresents both music and listeners' emotional experience by neglecting the role of time.

2.2 Musical Features

Most machine learning methods described in the literature use the audio to extract the musical features [7, 22, 10, 3]. Musical features such as tempo, loudness, and timbre, among many others, are estimated from the audio by means of signal processing algorithms [12]. Typically, these features are calculated from successive frames taken from excerpts of the audio that last a few seconds [7, 22, 10, 3, 4] and then averaged, losing the temporal correlation [10]. Consequently, the whole piece (or track) is represented by a static (non time-varying) vector, intrinsically assuming that musical experience is static and that the listener's emotional response can be estimated from the audio alone. The term 'semantic gap' has been coined to refer to perceived musical information that does not seem to be contained in the acoustic patterns present in the audio, even though listeners agree about its existence [21].

However, to fully understand emotional expression in music, it is important to study the performer's and composer's intention on the one hand, and the listener's perception on the other [6]. Music happens essentially in the brain, so we need to take the cognitive mechanisms involved in processing musical information into account if we want to be able to model people's emotional response to music. Low-level audio features give rise to high-level musical features in the brain, and these, in turn, influence emotion recognition (and experience). This is where we argue that time has a major role, still neglected in most approaches found in the literature. Musical experience and the cognitive processes that regulate musical emotions are entangled with each other around the temporal dimension, so the model of emotion should account for that.

2.3 Representation of Emotions

MER research tends to use categorical descriptions of emotions where the investigator selects a set of "emotional labels" (usually mutually exclusive). The left-hand side of figure 1 illustrates these emotional labels (Hevner's adjective circle [2]) clustered in eight classes. The choice of the emotional labels is important and might even affect the results. For example, the terms associated with music usually depend on genre (pop music is much more likely than classical music to be described as "cool"). As Yang [22] points out, the categorical representation of emotions faces a granularity issue because the number of classes might be too small to span the rich range of emotions perceived by humans. Increasing the number of classes does not necessarily solve the problem because the language used to categorize emotions is ambiguous and subjective [1]. Therefore, some authors [7, 22] have proposed to adopt a parametric model from psychology research [14] known as the circumplex model of affect (CMA). The CMA consists of two independent dimensions whose axes represent continuous values

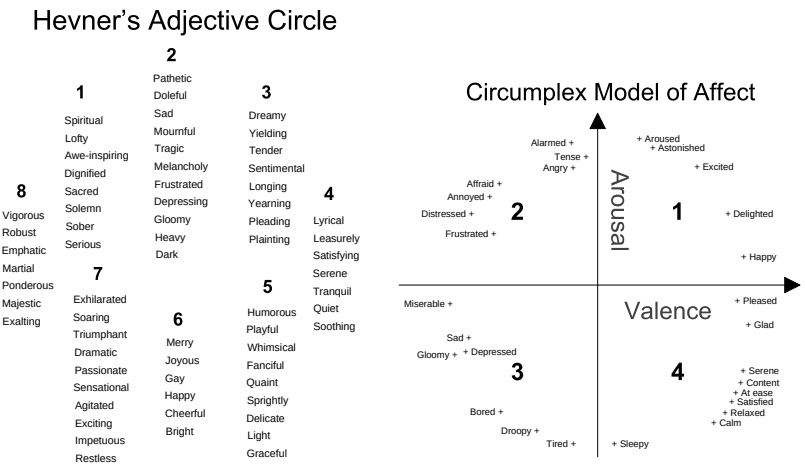


Fig. 1. Examples of models of emotion. The left-hand side shows Hevner's adjective circle [2], a categorical description. On the right, we see the circumplex model of affect [14], a parametric model.

of valence (positive or negative semantic meaning) and arousal (activity or excitation). The right-hand side of figure 1 shows the CMA and the position of some adjectives used to describe emotions associated with music in the plane. An interesting aspect of parametric representations such as the CMA lies in the continuous nature of the model and the possibility to pinpoint where specific emotions are located. Systems based on this approach train a model to compute the valence and arousal values and represent each music piece as a point in the two-dimensional emotion space [22].

One common criticism of the CMA is that the representation does not seem to be metric. That is, emotions that are very different in terms of semantic meaning (and psychological and cognitive mechanisms involved) can be close in the plane. In this article, we argue that the lack of temporal information is a much bigger problem because music happens over time and the way listeners associate emotions with music is intrinsically linked to the temporal evolution of the musical features. Also, emotions are dynamic and have distinctive temporal profiles (boredom is very different from astonishment in this respect, for example).

3 The Role of Time in the Complex Relationship Between Music and Emotions

Krumhansl [9] suggests that music is an important part of the link between emotions and cognition. More specifically, Krumhansl investigated how the dynamic aspect of musical emotion relates to the cognition of musical structure. According to Krumhansl, musical emotions change over time in intensity and quality,

and these emotional changes covary with changes in psycho-physiological measures [9]. Musical meaning and emotion depend on how the actual events in the music play against this background of expectations. David Huron [5] wrote that humans use a general principle in the cognitive system that regulates our expectations to make predictions. According to Huron, music (among other stimuli) influences this principle, modulating our emotions. Time is a very important aspect of musical cognitive processes. Music is intrinsically temporal and we need to take into account the role of human memory when experiencing music. In other words, musical experience is learned. As the music unfolds, the learned model is used to generate expectations, which are implicated in the experience of listening to music. Meyer [11] proposed that expectations play the central psychological role in musical emotions.

3.1 Temporal Evolution of Musical Features

The first important step to incorporate time into MER is to monitor the temporal evolution of musical features [18]. After the investigator chooses which features to use in a particular application, the feature vector should be calculated for every frame of the audio signal and kept as a time series (i.e., a time-varying vector of features). The temporal correlation of the features must be exploited and fed into the model of emotions to estimate listeners' response to the repetitions and the degree of "surprise" that certain elements might have [19].

Here we could make a distinction between perceptual features of musical sounds (such as pitch, timbre, and loudness) and musical parameters (such as tempo, key, and rhythm), related to the structure of the piece (and usually found in the score). Both of them contribute to listeners' perception of emotions. However, their temporal variations occur at different rates. Timbral variations, for example, and key modulations or tempo changes happen at different levels. Figure 2 illustrates these variations at the microstructural (musical sounds) and macrostructural (musical parameters) level.

3.2 Emotional Trajectories

A very simple way of recording information about the temporal variation of emotional perception of music would be to ask listeners to write down the emotional label and a time stamp as the music unfolds. The result is illustrated on the left-hand side of figure 3. However, this approach suffers from the granularity and ambiguity issues inherent of using a categorical description of emotions. Ideally, we would like to have an estimate of how much a certain emotion is present at a particular time.

Krumhansl [8] proposes to collect listener's responses continuously while the music is played, recognizing that retrospective judgments are not sensitive to unfolding processes. However, in this study [8], listeners assessed only one emotional dimension at a time. Each listener was instructed to adjust the position of a computer indicator to reflect how the amount of a specific emotion (for

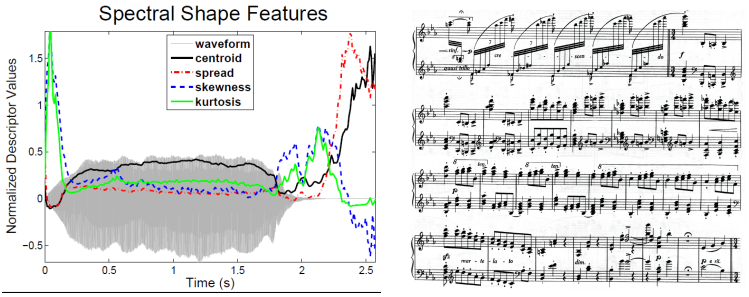


Fig. 2. Examples of the temporal variations of musical features. The left-hand side shows temporal variations of the four spectral shape features (centroid, spread, skewness, and kurtosis, perceptually correlated to timbre) during the course of a musical instrument sound (microstructural level). On the right, we see variations of musical parameters (macrostructural level) represented by the score for simplicity.

example, sadness) they perceived changed over time while listening to excerpts of pieces chosen to represent the emotions [8].

Here, we propose a similar procedure using a broader palette of emotions available to allow listeners to associate different emotions to the same piece. Recording listener's emotional ratings over time [13] would lead to an emotional trajectory like the one shown on the right of figure 3, which illustrates an emotional trajectory (time is represented by the arrow) in a conceptual emotional space, where the dimensions can be defined to suit the experimental setup. The investigator can choose to focus on specific emotions such as happiness and aggressiveness, for example. In this case, one dimension would range from happy to sad, while the other from aggressive to calm. However, we believe that Russell's CMA [14] would better fit the exploration of a broader range of emotions because the dimensions are not explicitly labeled as emotions.

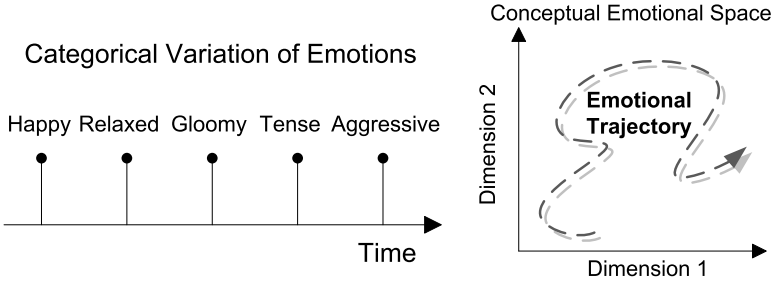


Fig. 3. Temporal variation of emotions. The left-hand side shows emotional labels recorded over time. On the right, we see a continuous conceptual emotional space with an emotional trajectory (time is represented by the arrow).

3.3 Investigating the Relationship Between the Temporal Evolution of Musical Features and the Emotional Trajectories

Finally, we should investigate the relationship between the temporal variation of musical features and the emotional trajectories. MER systems should include information about the rate of temporal change of musical features. For example, we should investigate how changes in loudness correlate with the expression of emotions. Schubert [18] studied the relationship between musical features and perceived emotion using continuous response methodology and time-series analysis. Musical features (loudness, tempo, melodic contour, texture, and spectral centroid) were differenced and used as predictors in linear regression models of valence and arousal. This study found that changes in loudness and tempo were associated positively with changes in arousal, and melodic contour varied positively with valence. When Schubert [19] discussed modeling emotion as a continuous, statistical function of musical parameters, he argued that the statistical modeling of memory is a significant step forward in understanding aesthetic responses to music. Only very recently MER systems started incorporating dynamic changes in efforts mainly by Schmidt and Kim [15–17, 20]. Therefore, this article aims at motivating the incorporation of time in MER to help break through the so-called "glass ceiling" (or "semantic gap") [21], improving the performance of computational models of musical emotion with advances in our understanding of the currently mysterious relationship between music and emotions.

4 Conclusions

Research on automatic recognition of emotion in music, still in its infancy, has focused on comparing "emotional labels" automatically calculated from different representations of music with those of human annotators. Usually the model represents the musical features as static vectors extracted from short excerpts and associates one emotion to each piece, neglecting the temporal nature of music. Studies in music psychology suggest that time is essential in emotional expression. In this article, we argue that MER systems must take musical context (what happened before) and listener expectations into account. We advocate the incorporation of time in both the representation of musical features and the model of emotions. We prompted MER researchers to represent the music as a time-varying vector of features and to investigate how the emotions evolve in time as the music develops, representing the listener's emotional response as an emotional trajectory. Finally, we discussed the relationship between the temporal evolution of the musical features and the emotional trajectories.

Future perspectives include the development of computational models that exploit the repetition of musical patterns and novel elements to predict listeners' expectations and compare them against the recorded emotional trajectories. Only by including temporal information in automatic recognition of emotions can we advance MER systems to cope with the complexity of human emotions in one of its canonical means of expression, music.

References

1. Gabrielsson, A., Lindstrom, E.: The Role of Structure in the Musical Expression of Emotions. In: Handbook of Music and Emotion: Theory, Research, Applications. Eds. Patrik N. Juslin and John Sloboda, pp. 367–400 (2011)
2. Hevner, K.: Experimental Studies of the Elements of Expression in Music. The Am. Journ. Psychology . 48 (2), pp. 246–268 (1936)
3. Hu, X., Downie, J.S., Laurier, C., Bay, M., and Ehmann, A.F.: The 2007 MIREX Audio Mood Classification Task: Lessons Learned. In: Proc. ISMIR (2008)
4. Huq, A., Bello, J.P., and Rowe, R.: Automated Music Emotion Recognition: A Systematic Evaluation. Journ. New Music Research. 39(4), pp. 227–244 (2010)
5. Huron, D.: Sweet Anticipation: Music and the Psychology of Expectation. MIT Press, (2006)
6. Juslin, P., Timmers, R.: Expression and Communication of Emotion in Music Performance. In: Handbook of Music and Emotion: Theory, Research, Applications Eds. Patrik N. Juslin and John Sloboda, pp. 453–489 (2011)
7. Kim, Y.E., Schmidt, E., Migneco, R., Morton, B., Richardson, P., Scott, J., Speck, J., Turnbull, D.: Music Emotion Recognition: A State of the Art Review. In: Proc. ISMIR (2010)
8. Krumhansl, C. L.: An Exploratory Study of Musical Emotions and Psychophysiology. Canadian Journ. Experimental Psychology. 51, pp. 336–352 (1997)
9. Krumhansl, C. L.: Music: A Link Between Cognition and Emotion. Current Directions in Psychological Science. 11, pp. 45–50 (2002)
10. MacDorman, K. F., Ough S., Ho C.C.: Automatic Emotion Prediction of Song Excerpts: Index Construction, Algorithm Design, and Empirical Comparison. Journ. New Music Research. 36, pp. 283–301 (2007)
11. Meyer, L.: Music, the Arts, and Ideas. University of Chicago Press, Chicago (1967)
12. Müller, M., Ellis, D.P.W., Klapuri, A., Richard, G.: Signal Processing for Music Analysis. IEEE Journal of Selected Topics in Sig. Proc. 5(6), pp. 1088–1110 (2011)
13. Nagel, F., Kopiez, R., Grewe, O., Altenmüller, E.: EMuJoy. Software for the Continuous Measurement of Emotions in Music. Behavior Research Methods, 39 (2), pp. 283–290 (2007)
14. Russell, J.A.: A Circumplex Model of Affect. Journ. Personality and Social Psychology. 39, pp. 1161–1178 (1980)
15. Schmidt, E.M., Kim, Y.E.: Prediction of Time-Varying Musical Mood Distributions Using Kalman Filtering. In: Proc. ICMLA (2010)
16. Schmidt, E.M., Kim, Y.E.: Prediction of Time-Varying Musical Mood Distributions from Audio. In: Proc. ISMIR (2010)
17. Schmidt, E.M., Kim, Y.E.: Modeling Musical Emotion Dynamics with Conditional Random Fields. In: Proc. ISMIR (2011)
18. Schubert, E.: Modeling Perceived Emotion with Continuous Musical Features. Music Perception, 21(4), pp. 561–585 (2004)
19. Schubert, E.: Analysis of Emotional Dimensions in Music Using Time Series Techniques. Journ. Music Research, 31, pp. 65–80 (2006)
20. Vaizman, Y., Granot, R.Y., Lanckriet, G.: Modeling Dynamic Patterns for Emotional Content in Music. In: Proc. ISMIR (2011)
21. Wiggins, G. A.: Semantic Gap?? Schemantic Schmap!! Methodological Considerations in the Scientific Study of Music. IEEE International Symposium on Multimedia, pp. 477-482 (2009)
22. Yang, Y., Chen, H.: Ranking-Based Emotion Recognition for Music Organization and Retrieval. IEEE Trans. Audio, Speech, Lang. Proc. 19, 4 (2011)

The Intervalgram: An Audio Feature for Large-scale Melody Recognition

Thomas C. Walters, David A. Ross, and Richard F. Lyon

Google, 1600 Amphitheatre Parkway, Mountain View, CA, 94043, USA
tomwalters@google.com

Abstract. We present a system for representing the melodic content of short pieces of audio using a novel chroma-based representation known as the 'intervalgram', which is a summary of the local pattern of musical intervals in a segment of music. The intervalgram is based on a chroma representation derived from the temporal profile of the stabilized auditory image [10] and is made locally pitch invariant by means of a 'soft' pitch transposition to a local reference. Intervalgrams are generated for a piece of music using multiple overlapping windows. These sets of intervalgrams are used as the basis of a system for detection of identical melodies across a database of music. Using a dynamic-programming approach for comparisons between a reference and the song database, performance is evaluated on the 'covers80' dataset [4]. A first test of an intervalgram-based system on this dataset yields a precision at top-1 of 53.8%, with an ROC curve that shows very high precision up to moderate recall, suggesting that the intervalgram is adept at identifying the easier-to-match cover songs in the dataset with high robustness. The intervalgram is designed to support locality-sensitive hashing, such that an index lookup from each single intervalgram feature has a moderate probability of retrieving a match, with few false matches. With this indexing approach, a large reference database can be quickly pruned before more detailed matching, as in previous content-identification systems.

Keywords: Melody Recognition, Auditory Image Model, Machine Hearing

1 Introduction

We are interested in solving the problem of cover song detection at very large scale. In particular, given a piece of audio, we wish to identify another piece of audio representing the same melody, from a potentially very large reference set. Though our approach aims at the large-scale problem, the representation developed is compared in this paper on a small-scale problem for which other results are available.

There can be many differences between performances with identical melodies. The performer may sing or play the melody at a different speed, in a different key or on a different instrument. However, these changes in performance do not, in general, prevent a human from identifying the same melody, or pattern of notes.

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

Thus, given a performance of a piece of music, we wish to find a representation that is to the largest extent possible invariant to such changes in instrumentation, key, and tempo.

Serra [12] gives a thorough overview of the existing work in the field of melody identification, and breaks down the problem of creating a system for identifying versions of a musical composition into a number of discrete steps. To go from audio signals for pieces of music to a similarity measure, the proposed process is:

- Feature extraction
- Key invariance (invariance to transposition)
- Tempo invariance (invariance to a faster or slower performance)
- Structure invariance (invariance to changes in long-term structure of a piece of music)
- Similarity computation

In this study, we concentrate on the first three of these steps: the extraction of an audio feature for a signal, the problem of invariance to pitch shift of the melody (both locally and globally) and the problem of invariance to changes in tempo between performances of a piece of music. For the first stage, we present a system for generating a pitch representation from an audio signal, using the stabilized auditory image (SAI) [10] as an alternative to standard spectrogram-based approaches. Key invariance is achieved locally (per feature), rather than globally (per song). Individual intervalgrams are key normalized relative to a reference chroma vector, but no guarantees are made that the reference chroma vector will be identical across consecutive features. This local pitch invariance allows for a feature that can track poor-quality performances in which, for example, a singer changes key gradually over the course of a song. It also allows the feature to be calculated in a streaming fashion, without having to wait to process all the audio for a song before making a decision on transposition. Other approaches to this problem have included shift-invariant transforms [9], the use of all possible transpositions [5] or finding the best transposition as a function of time in a symbolic system [13]. Finally, tempo invariance is achieved by the use of variable-length time bins to summarize both local and longer-term structure. This approach is in contrast to other systems [5, 9] which use explicit beat tracking to achieve tempo invariance.

While the features are designed for use in a large-scale retrieval system when coupled with a hashing technique [1], in this study we test the baseline performance of the features by using a Euclidean distance measure. A dynamic-programming alignment is performed to find the smallest-cost path through the map of distances between a probe song and a reference song; partial costs, averaged over good paths of reasonable duration, are used to compute a similarity score for a each probe-reference pair.

We evaluate performance of the intervalgam (using both SAI-based chroma and spectrogram-based chroma) using the 'covers80' dataset [4]. This is a set of 160 songs, in 80 pairs that share an underlying composition. There is no explicit notion of a 'cover' versus an 'original' in this set, just an 'A' version and

a 'B' version of a given composition, randomly selected. While it is a small corpus, several researchers have made use of this dataset for development of audio features, and report results on it. Ellis [5] reports performance in terms of absolute classification accuracy for the LabRosa 2006 and 2007 music information retrieval evaluation exchange (MIREX) competition, and these results are extended by, amongst others, Ravuri and Ellis [11], who present detection error tradeoff curves for a number of systems.

Since we are ultimately interested in the use of the intervalgram in a large-scale system, it is worth briefly considering the requirements of such a system. In order to perform completely automated detection of cover songs from a large reference collection, it is necessary to tune a system to have extremely low false hit rate on each reference. For such a system, we are interested less in high absolute recall and more in finding the best possible recall given a very low threshold for false positives. Such systems have previously been reported for nearly-exact-match content identification [1]. The intervalgram has been developed for and tested with a similar large-scale back end based on indexing, but there is no large accessible data set on which performance can be reported. It is hard to estimate recall on such undocumented data sets, but the system identifies a large number of covers even when tuned for less than 1% false matches.

2 Algorithm

2.1 The Stabilized Auditory Image

The stabilized auditory image (SAI) is a correlogram-like representation of the output of an auditory filterbank. In this implementation, a 64-channel pole-zero filter cascade [8] is used. The output of the filterbank is half-wave rectified and a process of 'strobe detection' is carried out. In this process, large peaks in the waveform in each channel are identified. The original waveform is then cross-correlated with a sparsified version of itself which is zero everywhere apart from at the identified strobe points. This process of 'strobed temporal integration' [10, 14] is very similar to performing autocorrelation in each channel, but is considerably cheaper to compute due to the sparsity of points in the strobe signal. The upper panels of Figure 1 show a waveform (upper panel) and stabilized auditory image (middle panel) for a sung note. The pitch of the voice is visible as a series of vertical ridges at lags corresponding to multiples of the repetition period of the waveform, and the formant structure is visible in the pattern of horizontal resonances following each large pulse.

2.2 Chroma From the Auditory Image

To generate a chroma representation from the SAI, the 'temporal profile' is first computed by summing over the frequency dimension; this gives a single vector of values which correspond to the strength of temporally-repeating patterns in the waveform at different lags. The temporal profile gives a representation of

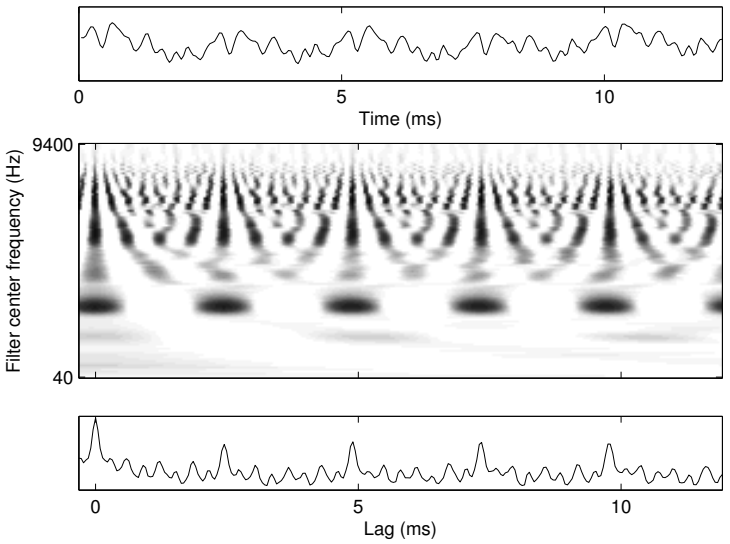


Fig. 1. Waveform (top panel), stabilized auditory image(SAI) (middle panel) and SAI temporal profile (bottom panel) for a human voice singing a note.

the time intervals associated with strong temporal repetition rates, or possible pitches, in the incoming waveform. This SAI temporal profile closely models human pitch perception [6]; for example, in the case of stimuli with a missing fundamental, there may be no energy in the spectrogram at the frequency of the pitch perceived by a human, but the temporal profile will show a peak at the time interval associated with the missing fundamental.

The lower panel of Figure 1 shows the temporal profile of the stabilized auditory image for a sung vowel. The pitch is visible as a set of strong peaks at lags corresponding to integer multiples of the pulse rate of the waveform. Figure 2 shows a series of temporal profiles stacked in time, a 'pitch-o-gram', for a piece of music with a strong singing voice in the foreground. The dark areas correspond to lags associated with strong repetition rates in the signal, and the evolving melody is visible as a sequence of horizontal stripes corresponding to notes; for example in the first second of the clip there are four strong notes, followed by a break of around 1 second during which there are some weaker note onsets.

The temporal profile is then processed to map lag values to pitch chromas in a set of discrete bins, to yield a representation as chroma vectors, also known as 'pitch class profiles' (PCPs) [12]. In our standard implementation, we use 32 pitch bins per octave. Having more bins than the standard 12 semitones in the Western scale allows the final feature to accurately track the pitch in recordings where

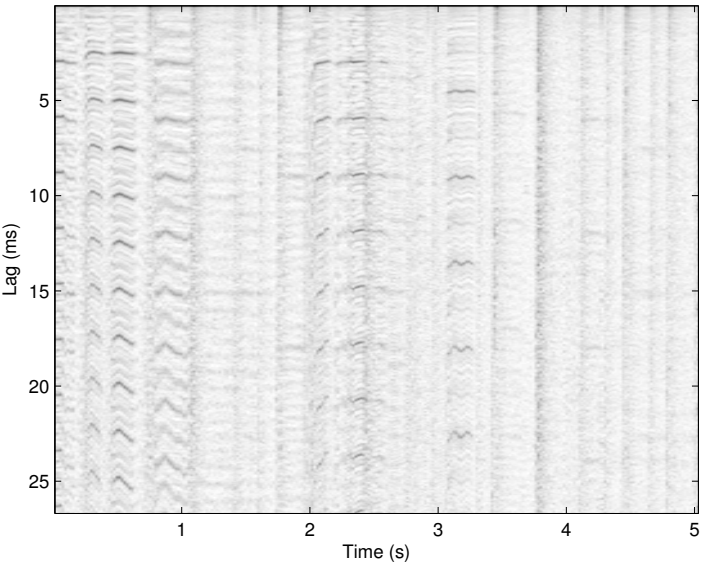


Fig. 2. A 'pitch-o-gram' created by stacking a number of SAI temporal profiles in time. The lag dimension of the auditory image is now on the vertical axis. Dark ridges are associated with strong repetition rates in the signal.

the performer is either mistuned or changes key gradually over the course of the performance; it also enables more accurate tracking of pitch sweeps, vibrato, and other non-quantized changes in pitch. Additionally, using an integer power of two for the dimensions of the final representation lends itself to easy use of a wavelet decomposition for hashing, which is discussed below. The chroma bin assignment is done using a weighting matrix, by which the temporal profile is multiplied to map individual samples from the lag dimension of the temporal profile into chroma bins. The weighting matrix is designed to map the linear time-interval axis to a wrapped logarithmic note pitch axis, and to provide a smooth transition between chroma bins. An example weighting matrix is shown in Figure 3. The chroma vectors for the same piece of music as in Figure 2 are shown in Figure 4.

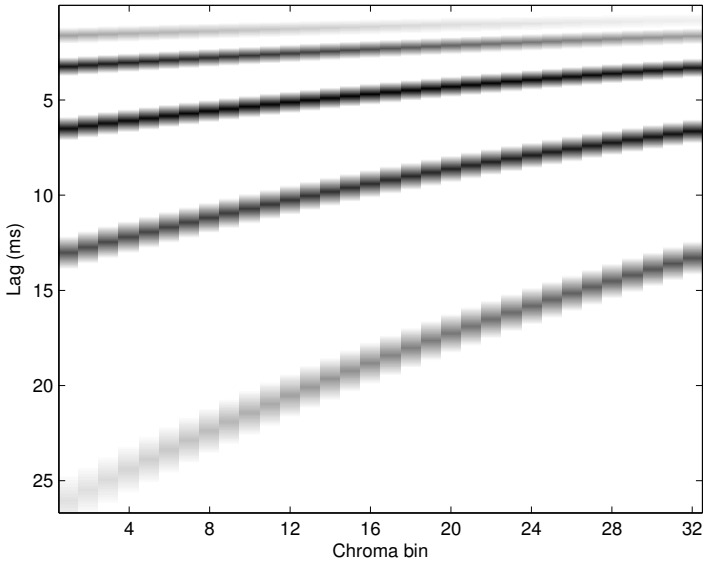


Fig. 3. Weighting matrix to map from the time-lag axis of the SAI to chroma bins.

2.3 Chroma From the Spectrogram

In addition to the SAI-based chroma representation described above, a more standard spectrogram-based chroma representation was tested as the basis for the intervalgram. In this case, chroma vectors were generated using the `chromagram_E` function distributed with the covers80 [4] dataset, with a modified step size to generate chroma vectors at the rate of 50 per second, and 32 pitch bins per

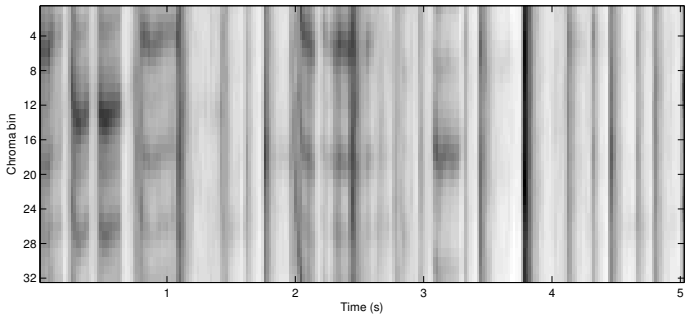


Fig. 4. Chroma vectors generated from the pitch-o-gram vectors shown in Figure 2.

octave for compatibility with the SAI-based features above. This function uses a Gaussian weighting function to map FFT bins to chroma, and weights the entire spectrum with a Gaussian weighting function to emphasize octaves in the middle of the range of musical pitches.

2.4 Intervalgram Generation

A stream of chroma vectors is generated at a rate of 50 per second. From this chromagram, a stream of 'intervalgrams' is constructed at the rate of around 4 per second. The intervalgram is a matrix with dimensions of chroma and time offset; however, depending on the exact design the time-offset axis may be nonlinear.

For each time-offset bin in the intervalgram, a sequence of individual chroma vectors are averaged together to summarize the chroma in some time window, before or after a central reference time. It takes several contiguous notes to effectively discern the structure of a melody, and for any given melody the stream of notes may be played a range of speeds. In order to take into account both short- and longer-term structure in the melody, a variable-length time-averaging process is used to provide a fine-grained view of the local melody structure, and simultaneously give a coarser view of longer timescales, to accommodate a moderate amount of tempo variation; that is, small absolute time offsets use narrow time bin widths, while larger absolute offsets use larger bin widths. Figure 5 shows how chroma vectors are averaged together to make the intervalgram. In the examples below, the widths of the bins increase from the center of the intervalgram, and are proportional to the sum of a forward and reverse exponential $w_b = f\left(w_f^p + w_f^{-p}\right)$, where p is an integer between 0 and 15 (for the positive bins) and between 0 and -15 (for the negative bins), f is the central bin width, and w_f is the width factor which determines the speed with which the bin width increases as a function of distance from the center of the intervalgram.

In the best-performing implementation, the temporal axis of the intervalgram is 32 bins wide and spans a total time window of around 30 seconds. The central

two slices along the time axis of the intervalgram are the average of 18 chroma vectors each (360ms each), moving away from the centre of the intervalgram, the outer temporal bins summarize longer time-scales before and after the central time. The number of chroma vectors averaged in each bin increases up to 99 (1.98s) in the outermost bins leading to a total temporal span of 26 seconds for each intervalgram.

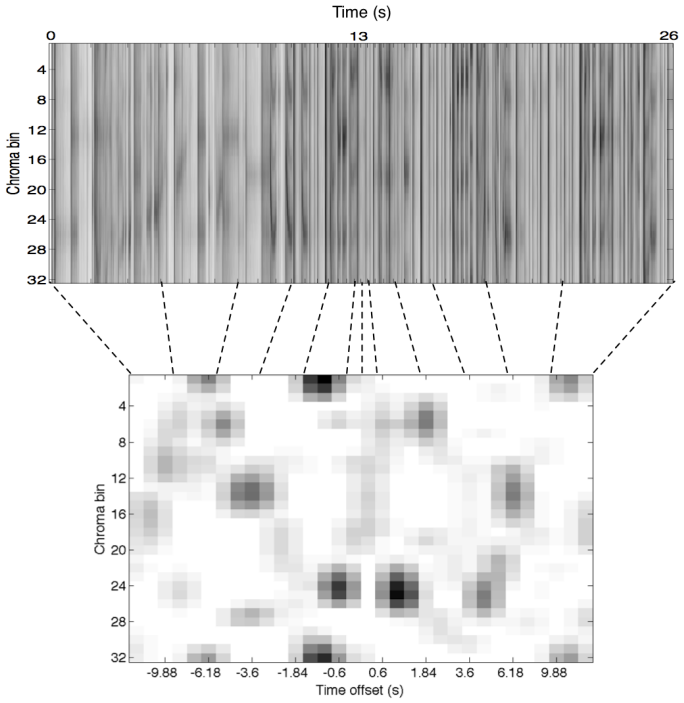


Fig. 5. The intervalgram is generated from the chromagram using variable-width time bins and cross-correlation with a reference chroma vector to normalize chroma within the individual intervalgram.

A 'reference' chroma vector is also generated from the stream of incoming chroma vectors at the same rate as the intervalgrams. The reference chroma vector is computed by averaging together nine adjacent chroma vectors using a triangular window. The temporal center of the reference chroma vector corresponds to the temporal center of the intervalgram. In order to achieve local pitch invariance, this reference vector is then circularly cross-correlated with each of the surrounding intervalgram bins. This cross-correlation process implements a 'soft' normalization of the surrounding chroma vectors to a prominent pitch or pitches in the reference chroma vector. Given a single pitch peak in the refer-

ence chroma vector, the process corresponds exactly to a simple transposition of all chroma vectors to be relative to the single pitch peak. In the case where there are multiple strong peaks in the reference chroma vector, the process corresponds to a simultaneous shifting to multiple reference pitches, followed by a weighted average based on the individual pitch strengths. This process leads to a blurry and more ambiguous interval representation but, crucially, never leads to a hard decision being made about the 'correct' pitch of the melody at any point. Making only 'soft' decisions at each stage means that there is less need for either heuristics or tuning of parameters in building the system. With standard parameters the intervalgram is a 32 by 32 pixel feature vector generated at the rate of one every 240ms and spanning a 26 second window. Since there are many overlapping intervalgrams generated, there are many different pitch reference slices used, some making crisp intervalgrams, and some making fuzzy intervalgrams.

2.5 Similarity Scoring

Dynamic programming is a standard approach for aligning two audio representations, and has been used for version identification by many authors (for example [16]; Serra [12] provides a representative list of example implementations). To compare sets of features from two recordings, each feature vector from the probe recording is compared to each feature vector from the reference recording, using some distance measure, for example Euclidean distance, correlation, or Hamming distance over a locality-sensitive hash of the feature. This comparison yields a distance matrix with samples from the probe on one axis and samples from the reference on the other. We then find a minimum-cost path through this matrix using a dynamic programming algorithm that is configured to allow jumping over poorly-matching pairs. Starting at the corner corresponding to the beginning of the two recordings the path can continue by jumping forward a certain number of pixels in both the horizontal and vertical dimensions. The total cost for any particular jump is a function of the similarity of the two samples to be jumped to, the cost of the jump direction and the cost of the jump distance. If two versions are exactly time-aligned, we would expect that the minimum-cost path through the distance matrix would be a straight line along the leading diagonal. Since we expect the probe and reference to be roughly aligned, the cost of a diagonal jump is set to be smaller than the cost of an off-diagonal jump.

The minimum and maximum allowed jump lengths in samples can be selected to allow the algorithm to find similar intervalgrams that are more sparsely distributed, interleaved with poorly matching ones, and to constrain the maximum and minimum deviation from the leading diagonal. Values that work well are a minimum jump of 3 and maximum of 4, with a cost factor equal to the longer of the jump dimensions (so a move of 3 steps in the reference and 4 in the probe costs as much as 4,4 even though it uses up less reference time, while jumps of 3,3 and 4,4 along the diagonal can be freely intermixed without affecting the score as long as enough good matching pairs are found to jump between). These lengths, along with the cost penalty for an off-diagonal jump and the difference

in cost for long jumps over short jumps, are parameters of the algorithm. Figure 6 shows a distance matrix for a probe and reference pair.

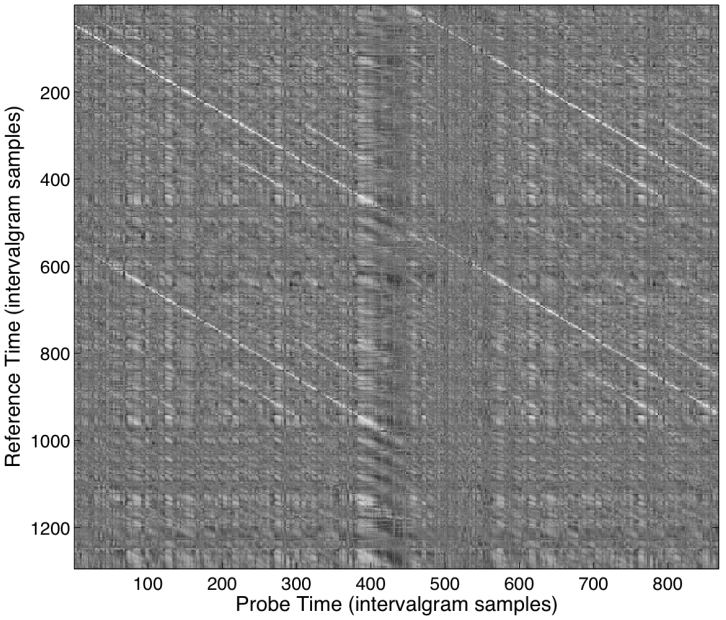


Fig. 6. Example distance matrix for a pair of songs which share an underlying melody. The lighter pixels show the regions where the intervalgrams match closely.

In the following section we test the performance of the raw intervalgrams, combined with the dynamic programming approach described above, in finding similarity between cover songs.

3 Experiments

We tested performance of the similarity-scoring system based on the intervalgram, as described above, using the standard paradigm for the covers80 dataset, which is to compute a distance matrix for all query songs against all reference songs, and report the percentage of query songs for which the correct reference song has the highest similarity score.

Intervalgrams were computed from the SAI using the parameters outlined in Table 1, and scoring of probe-reference pairs was performed using the dynamic programming approach described above. Figure 7 shows the matrix of scores for the comparison of each probe with all reference tracks. Darker pixels denote

lower score, and lighter pixels denote higher scores. The white crosses show the highest-scoring reference for a given probe. 43 of the 80 probe tracks in the covers80 dataset were correctly matched to their associated reference track leading to a score of 53.8% on the dataset. For comparison, Ellis [5] reports a score of 42.5% for his MIREX2006 entry, and 67.5% for his MIREX2007 entry (the latter had the advantage of using covers80 as a development set, so is less directly comparable).

Parameter	Value
Chromagram step size (ms)	20
Chroma bins per octave	32
Total intervalgram width (s)	26.04
Intervalgram step size (ms)	240
Reference chroma vector width (chroma vectors)	4

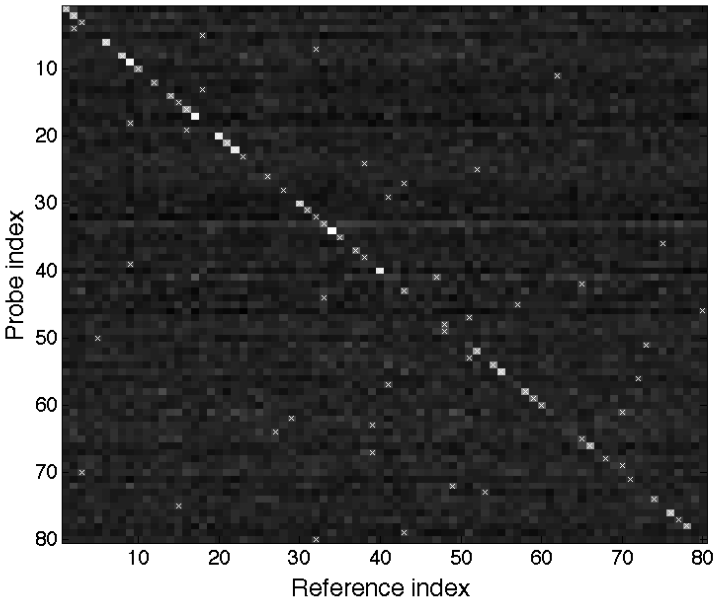


Fig. 7. Scores matrix for comparing all probes and references in the 'covers80' dataset. Lighter pixels denote higher scores, indicating a more likely match. White crosses denote the best-matching reference for each probe.

In addition to the SAI-based chroma features, standard spectrogram-based chroma features were computed from all tracks in the 'covers80' dataset. These features used 32 chroma bins, and were computed at 50 frames per second, to

provide a drop-in replacement for the SAI-based features. Intervalgrams were computed from these features using the parameters in Table 1.

In order to generate detection error tradeoff curves for the dataset, the scores matrix from Figure 7 was dynamically thresholded to determine the number of true and false positives for a given threshold level. The results were compared against the reference system supplied with the covers80 dataset, which is essentially the same as the system entered by LabRosa for the 2006 MIREX competition, as documented by Ellis [5]. Figure 8 shows ROC curves the Elllis MIREX'06 entry and for the intervalgram-based system, both with SAI chroma features and spectrogram chroma features. Re-plotting the ROC curve as a DET curve to compare results with Ravuri and Ellis [11], performance of the intervalgram-based features is seen to consistently lie between that of the LabRosa MIREX 2006 entry and their 2007 entry.

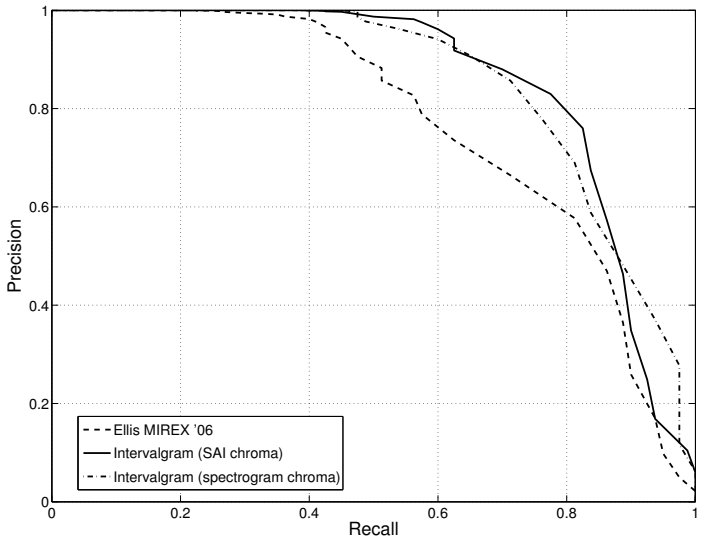


Fig. 8. ROC curves for the intervalgram-based system described in this paper and the LabROSA MIREX 2006 entry [5].

Of particular interest is the performance of the features at high precision. The SAI-based intervalgram can achieve 47.5% recall at 99% precision, whereas the Ellis MIREX '06 system achieves 35% recall at 99% precision. These early results suggest that the intervalgram shows good robustness to interference. The intervalgram also stands up well to testing on larger, internal, datasets in combination with hashing techniques, as discussed below.

4 Discussion

We have introduced a new chroma-based feature for summarizing musical melodies, which does not require either beat tracking or exhaustive search for transposition invariance, and have demonstrated a good baseline performance on a standard dataset. However, we developed the intervalgram representation to be a suitable candidate for large-scale, highly robust cover-song detection. In the following sections we discuss some approaches to the application of the intervalgram in such a system.

4.1 SAI and Spectrogram-based Chroma

There was no great difference in performance between intervalgrams generated using the temporal profile of the SAI and intervalgrams generated using a spectrogram-based chroma feature. However, there are some small differences in different regions of the ROC curve. Recall at high precision is very similar for both forms of chroma features; as precision is allowed to fall, the SAI-based features lead to slightly higher recall for a given precision, but the trend is reversed in the lower-precision end of the curve. This may suggest that there would be a benefit in combining both SAI-based and spectrogram-based chroma into a feature which makes use of both. There is some evidence to suggest that the temporal profile of the SAI may be robust to stimuli in which the pitch is ambiguous [6], but this result may be less relevant in the context of music.

4.2 Scaling Up

In order to perform melody recognition on a large database of content, it is necessary to find a cheaper and more efficient way of matching a probe song against many references. The brute-force approach of computing a full distance map for the probe against every possible reference scales as the product of the number of probes and the number of references; thus a system which makes it cheap to find a set of matching segments in all references for a given probe would be of great value. Bertin-Mahieux and Ellis [2] presented a system using hashed chroma landmarks as keys for a linear-time database lookup. Their system showed promise, and demonstrated a possible approach to large-scale cover-song detection but the reported performance numbers would not make for a practically-viable system. While landmark or 'interest point' detection has been extremely successful in the context of exact audio matching in noise [15] its effectiveness in such applications is largely due to the absolute invariance in the location of strong peaks in the spectrogram. For cover version identification the variability in performances, both in timing and in pitch, means that descriptors summarizing small constellations of interest points will necessarily be less discriminative than descriptors summarizing more complete features over a long time span. With this in mind, we now explore some options for generating compact hashes of full intervalgrams for indexing and retrieval purposes.

Hashing of the Intervalgram Using the process outlined above, 32×32 pixel intervalgrams are generated from a signal at the rate of one per 240ms. To effectively find alternative performances of a melody in a large-scale database, it must be possible to do efficient lookup to find sequences of potentially potential matching intervalgrams. The use of locality-sensitive-hashing (LSH) techniques over long-timescale features for music information retrieval has previously been investigated and found to be useul for large datasets [3]. Various techniques based on locality-sensitive hashing (LSH) may be employed to generate a set of compact hashes which summarize the intervalgram, and which can be used as keys to look up likely matches in a key-value lookup system.

An effective technique for summarizing small images with a combination of wavelet analysis and Min-Hash was presented by Baluja and Covell [1] in the context of hashing spectrograms for exact audio matching. A similar system of wavelet decomposition was previously applied to image analysis [7]. The system described in [1] has been adapted to produce a compact locality-sensitive hash of the intervalgram. The 32×32 intervalgram is decomposed into a set of wavelet coefficients using a Haar kernel, and the top t wavelet coefficients with the highest magnitude values retained. If the value t is chosen to be much smaller than the total number of pixels in the image, the most prominent structure of the intervalgram will be maintained, with a loss of some detail.

Compared to exact-match audio identification, this system is much more challenging, since the individual hash codes are noisier and less discriminative. The indexing stage necessarily has many false hits when it is tuned to get any reasonable recall, so there are still many (at least thousands out of a reference set of millions) of potential matches to score in detail before deciding whether there is a match.

5 Conclusions

The intervalgram is a pitch-shift-independent feature for melody-recognition tasks. Like other features for melody recognition, it is based on chroma features, but in our work the chroma representation is derived from the temporal profile of a stabilized auditory image, rather than from a spectrogram. To achieve pitch-shift invariance, individual intervalgrams are shifted relative to a reference chroma vector, but no global shift invariance is used. Finally, to achieve some degree of tempo-invariance, variable-width time-offset bins are used to capture both local and longer-term features.

In this study, the performance of the intervalgram was tested by using dynamic-programming techniques to find the cheapest path through similarity matrices comparing a cover song to all references in the 'covers80' dataset. Intervalgrams, followed by dynamic-programming alignment and scoring, gave a precision at top-1 of 53.8%. This performance value, and the associated ROC curve, lies between the performance of the Ellis 2006 and Ellis 2007 MIREX entries (the latter of which was developed using the covers80 dataset).

The intervalgram has shown itself to be a promising feature for melody recognition. It has good performance characteristics for high-precision matching with a low false-positive rate. Furthermore the algorithm is fairly simple and fully 'feed-forward', with no need for beat tracking or computation of global statistics. This means that it can be run in a streaming fashion, requiring only buffering of enough data to produce the first intervalgram before a stream of intervalgrams can be generated. This feature could make it suitable for applications like query-by-example in which absolute latency is an important factor.

We believe that the intervalgram representation would also lend itself well to large scale application when coupled with locality-sensitive hashing techniques such as wavelet-decomposition followed by minhash. The high precision would allow for querying of a large database with a low false-positive rate, and indeed preliminary experiments show some promise in this area. We look forward to tuning the performance of the intervalgram representation on larger research datasets.

References

1. S. Baluja and M. Covell. Waveprint: Efficient wavelet-based audio fingerprinting. *Pattern recognition*, 41(11):3467–3480, 2008.
2. T. Bertin-Mahieux and D. Ellis. Large-scale cover song recognition using hashed chroma landmarks. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, 2011.
3. M. Casey, C. Rhodes, and M. Slaney. Analysis of minimum distances in high-dimensional musical spaces. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(5):1015–1028, 2008.
4. D. Ellis, The 'covers80' cover song data set. http://labrosa.ee.columbia.edu/projects/coversongs/covers80/.
5. D. Ellis and C. Cotton. The 2007 LabROSA cover song detection system. *MIREX 2007 Audio Cover Song Evaluation system description*, 2007.
6. D. Ives and R. Patterson. Pitch strength decreases as f0 and harmonic resolution increase in complex tones composed exclusively of high harmonics. *The Journal of the Acoustical Society of America*, 123:2670, 2008.
7. C. Jacobs, A. Finkelstein, and D. Salesin. Fast multiresolution image querying. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 277–286. ACM, 1995.
8. R. Lyon. Cascades of two-pole-two-zero asymmetric resonators are good models of peripheral auditory function. *Journal of the Acoustical Society of America*, 130(6):3893, 2011.
9. M. Marolt. A mid-level representation for melody-based retrieval in audio collections. *Multimedia, IEEE Transactions on*, 10(8):1617–1625, 2008.
10. R. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand. Complex sounds and auditory images. In *Auditory physiology and perception, Proceedings of the 9th International Symposium on Hearing*, pages 429–446. Pergamon, 1992.
11. S. Ravuri and D. Ellis. Cover song detection: from high scores to general classification. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 65–68. IEEE, 2010.

12. J. Serra Julia. *Identification of versions of the same musical composition by processing audio descriptions*. PhD thesis, Universitat Pompeu Fabra, 2011.
13. W. Tsai, H. Yu, and H. Wang. Using the similarity of main melodies to identify cover versions of popular songs for music document retrieval. *Journal of Information Science and Engineering*, 24(6):1669–1687, 2008.
14. T. Walters. *Auditory-based processing of communication sounds*. PhD thesis, University of Cambridge, 2011.
15. A. Wang. An industrial strength audio search algorithm. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, volume 2, 2003.
16. C. Yang. Music database retrieval based on spectral similarity. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, 2001.

Perceptual dimensions of short audio clips and corresponding timbre features

Jason Jiří Musil, Budr Elnusairi, and Daniel Müllensiefen

Goldsmiths, University of London
j.musil@gold.ac.uk

Abstract. This study applied a multi-dimensional scaling approach to isolating a number of perceptual dimensions from a dataset of human similarity judgements for 800ms excerpts of recorded popular music. These dimensions were mapped onto the 12 timbral coefficients from the Echo Nest's Analyzer. Two dimensions were identified by distinct coefficients, however a third dimension could not be mapped and may represent a musical feature other than timbre. Implications are discussed within the context of existing research into human musical cognition. Suggestions for further research are given, which may help to establish whether surface features are processed using a common feature set (as in many music information retrieval systems), or whether individuals use features idiosyncratically to quickly process surface features of music.

Keywords: Timbre perception, short audio clips, similarity perception, sorting paradigm, MDS

1 Introduction

Many application systems in music information retrieval rely on some kind of timbre representation of music [1, 2]. Timbre, or the surface quality of sound, seems to be a core aspect of computational systems which compare, classify, organise, search, and retrieve music. This dominance of timbre and sound representations in modern user-targeted audio application systems might be partially explained by the importance of the perceptual qualities of sound in popular music; writing about pop music in 1987, sociomusicologist Simon Frith already noted that "The interest today (...) is in constantly dealing with new textures" [3]. Whilst musical textures can contain a lot of musical structure, they also depend on surface features separate from any musical syntax or structure, such as the harmonicity of sound, the timbral and acoustical qualities of instruments and spaces, and recording or post-production methods. The precision with which many features of sound can be defined and implemented through modern signal processing has surely also contributed to their popularity in the information retrieval community. Acoustic and timbral features have been defined as part of the MPEG4 and MPEG7 standards and are easily implemented where not already available from one of many software libraries.

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

Timbral features are popular in research and commercial music retrieval applications, yet there is surprisingly little rigorous research into perceptual principles explaining how certain timbral features can deliver results which are largely compatible with human music processing. Psychological and perceptual discourse around auditory processing often seems to be out of touch with parts of the audio engineering community. For example, an oft-cited validation of mel-frequency cepstral coefficients (MFCCs) as corresponding to human perceptual processing of sound is a brief engineering paper, rather than a psychological or psycho-acoustical study [4]. Conversely, some studies of human timbre perception (e.g. [5]) may have been unfairly overlooked by the psychological music research community due to their use of 'artificial' stimuli. Also, psychological studies of musical timbre have traditionally focused on the acoustics of musical instruments, or timbral qualities imparted by individual performers (e.g. vibrato, alteration of instrumental attack and decay). These are often studied in isolation and usually with reference to styles of Western art music (e.g. [6]; see [7] for an overview). Thus there is something of a discrepancy between the scope of psychological inquiries and the broader, data-driven goals of music information retrieval (MIR) as applied to finished recordings of popular music. This may exacerbate the relative ignorance between both fields.

The current study aims to bridge this gap to some extent by presenting data from a psychological experiment on human perception of timbral similarity, using short excerpts of Western commercial pop music as stimuli. In addition, this study also tries to identify the perceptual dimensions that Western listeners use when making similarity judgements based on timbre cues and to relate these to a set of timbral features that are well known to both music information researchers and software engineers: the 12 timbre feature coefficients provided through the Echo Nest Analyzer API[1]. As these involve considerable auditory modeling and dimensional reduction motivated to approximate human perception [8], we assume that the human and machine feature extractors under comparison are at least notionally parallel processes.

In this study, participants listen to very short excerpts of recorded commercial popular music and sort them into homogeneous groups. The paradigm is inspired by recent studies on genre [9] and song identification [10], which demonstrated that listeners are able to perform highly demanding tasks on the basis of musical information that is present in sub-second audio clips. Gjerdingen and Perrott found that 44% of participants' genre classifications of 250ms excerpts of commercially available music agreed with classifications they made of the same extracts when they were played for 3 seconds [9]. Krumhansl found that listeners could even identify the artists and titles of 25% of a series of 400ms clips of popular music spanning four decades [10]. At this timescale there are few, if any discernible melodic, rhythmic, harmonic or metric relationships to base judgements on. When musical-structural information is minimal, timbral information can be high; task performance also increased monotonically with longer exposures in both of the aforementioned studies.

[1] `http://developer.echonest.com/`

Many kinds of timbral information can be extracted from musical excerpts. The presence of typical instrumental sounds can undoubtedly help to identify a particular genre [9] and perception of key spectral and dynamic features is robust even for incomplete instrumental tones [11]. However, if timbre is defined more broadly as the spectro-temporal quality of sound, many surface features of polyphonic music could potentially be seen as coefficients in a timbre space. Indeed, the expression of musical emotion can be ascertained from 250ms of exposure, and familiarity with a piece from 500ms [12]. Spectral coefficients also join metric cues as predictors of surface judgements of musical complexity [13]. Different recording and production techniques can give rise to a plethora of perceptual timbral dimensions [14, 15].

In this study, in order to establish how non-expert listeners make use of musical surface features in a similarity sorting task we first apply multi-dimensional scaling (MDS) to extract a small number of perceptual dimensions and then relate these to coefficients in a timbre space. The timbral coefficients returned by the Echo Nest's Analyze service were chosen as the initial pool, as they have been usefully applied in a number of real-world applications. This research paradigm was established by classic studies on timbral perceptual dimensions for instrumental tones [16, 17], and is sensitive to subtle processing differences not picked up by traditional discrimination paradigms [18].

2 Method

131 participants (59 male, with a mean age of 30.8, SD=11.8) sorted 16 randomly ordered excerpt test-items into four equally sized bins. Sorts were unconstrained (other than the need for solutions to have exactly four items per bin) and participants could audition items at will. The set contained four each of jazz, rock, pop and hip-hop items, taken from songs identified on the `http://www.allmusic.com` website as being genre-typical but not universally known (i.e. through not having achieved the highest pop chart ratings). Genres were chosen on the basis of Rentfrow and Gosling's high-level categories of musical genre: reflective/complex (jazz), energetic/rhythmic (hip-hop), upbeat/conventional (pop), and intense/aggressive (rock) [19]. Genre-category ratings for these are stable over time and appear to correlate somewhat with stable personality traits [20]. Participants could thus solve the task implicitly (by perceived similarity) even if they possessed no genre-specific knowledge. By focusing on these categories, we also avoided the inherent instability and fluidity of industry genre boundaries. Gjerdingen and Perrott also found that the presence of vocals in extracts reduced genre rating performance [9]. Although vocal features are important for recognising musical styles (and this is reflected in the technologies used in MIR) we chose stimuli without vocals to avoid making the already short excerpts too difficult to classify. Excerpts were representative of the typical instrumentation of the song. Several sets were tested, however results from only one of the 800ms item-sets are analysed here, following piloting which suggested this set to

have desirable psychometric properties[2]. Vectors of timbral features for the same items were extracted through the Echo Nest's Analyzer and used as predictors of item-placement on these dimensions.

3 Analyses and Results

Each possible pair of clips received a score based on the number of participants assigning both clips in the pair to the same group. The resulting distance matrix was taken as an input to the non-metric multi-dimensional scaling procedure as implemented in the `R`-function `isoMDS` (from package `MASS`). Computing a 2- as well as a 3-dimensional solution we obtained stress values of 12.05 and 6.52 respectively, indicating a much better fit of the 3-dimensional solution to the data, with the 3-dimensional solution also satisfying the elbow criterion in a stress plot (not reproduced here). As a rule of thumb, Kruskal considers MDS solutions with a stress of 5 or lower a good fit while solutions with a stress value of 10 are still fair [21]. Thus, it seems that 3 dimensions are sufficient to describe the participants' perceptual judgements. The 3-dimensional solution is shown in *Figure 1.* Clustering of clips by genre in the MDS space is clearly visible.

As a subsequent step we tried to identify the 3 perceptual dimensions identified by MDS with any of the Echo Nest's 12 timbre coefficients. The Echo Nest Analyzer divides audio into segments with stable tonal content, i.e. roughly per note or chord. For each audio clip we obtained 2 to 5 segments with 12 timbre coefficients each. In order to obtain a homogeneous set of timbral features to compare to the 3 MDS dimensions we used a simple first-order linear model of the time series values of each coefficient for each clip. From each linear model we used the intercept (mean value) and the variance across the number of segments as an indicator of variability of the coefficient in the given clip. In addition, we used the number of segments per coefficient and clip as another indicator of tonal variability.

The pair-wise distributions and correlations between each MDS-dimension and the means and variances of the 12 coefficients indicated that the relationships between the perceptual dimensions and the timbral coefficients are mainly non-linear and distributions are far from normal. We therefore chose a random forest as an analysis technique, as it is able to model non-linear relationships and can additionally deal with a relatively high number of predictors (means and variances for each of the 12 coefficients plus the number of segments resulted in 25 predictor variables) compared to the low number of observations (16 audio clips; for a discussion of random forests as a classification and regression technique see chapter 15 in [22]). More specifically, we chose the conditional random forest model as implemented in the `R` package `party` [23], which is assumed to deliver more reliable estimates of variable importance when predictors are highly correlated and represent different measurement levels [24].

[2] A floor effect for 400ms stimuli was significantly less pronounced for 800ms stimuli in a pilot dataset with 117 participants (800ms per-item successful pairs out of a maximum of 3: M=1.22, SD=0.44; 400ms: M=1.05, SD=0.37; $t_{(31)}$=4.87, p<.001).

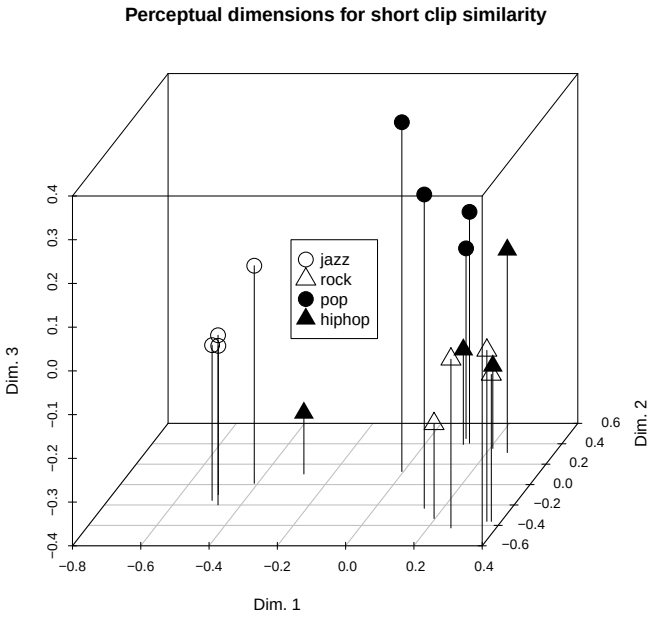


Fig. 1. The 3-dimensional solution of pairwise item distances. Points are differentiated by genre.

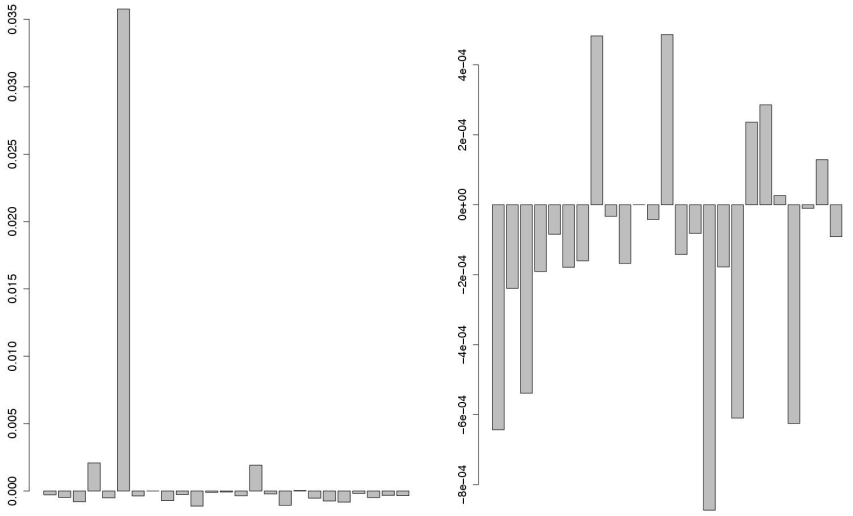


Fig. 2. Predictor importance for perceptual similarity dimensions 1 (*left*) and 3 (*right*). The tall bar for dimension 1 is the intercept of timbre coefficient 5. Note that the plots do not share a common y-axis.

Fitting a random forest model yielded a list of variable-importance values based on the usefulness of individual predictors for accurately predicting the so-called 'out-of-the-bag' (i.e. cross-validation) sample. The intercept (i.e. the mean) of the Echo Nest's timbral coefficient 5 was found to be of high importance as a predictor of perceptual dimension 1. A similarly clear picture was found for the intercept of coefficient 9, being highly important as a predictor of perceptual dimension 2. However, the picture was less clear for perceptual dimension 3, where all importance values for all variables remained within the margin of error around 0, indicating that perceptual dimension 3 cannot be closely associated with any (studied) timbral coefficient. Importance values of variables based on timbre coefficients are given in *Figure 2* for dimensions 1 and 3 for comparison.

4 Discussion

Three perceptual dimensions explained listeners' similarity judgements of short musical clips. Two of these dimensions were predicted by distinct surface features. Mean values but not variances of coefficients were selected as important predictors, which is interesting because the excerpts were long enough to contain some note- and beat-like temporal variations. Unfortunately, only a few timbral features returned by the Echo Nest are publicly documented, so it is difficult to say what these correspond to. A scale-less spectrogram in the existing documentation[3] suggests that coefficient 5, which predicted the coordinates of the 16 clips in perceptual dimension 1, might be a kind of mid-range filter. This would not be surprising, as spectral and dynamic effects are used to add low-end power and high-end presence to recordings. This could reduce the amount of useful information contained in those frequency bands, whilst the mid-range could become the most informative for clip discrimination and classification. Indeed, the most distant cluster on this dimension was jazz, which tends towards conservative mastering and emphasises distinctive instrumental timbres.

The distribution of clips along perceptual dimension 2, as well as incomplete information from the Echo Nest documentation for coefficient 9, suggested that this dimension may represent a similar filtering function to coefficient 5, albeit shifted higher or polarised more to high and low frequency bands. Despite this evidence for possible commonality between the human and machine feature extractors under study, dimension 3 is not predicted by any of the 12 Echo Nest timbral coefficients. At 800ms, the stimuli we used contain rudimentary information about tempo, chord changes, and rhythm. It is possible that dimension 3 represents the influence of such abstracted structures. The results obtained from studies with shorter stimuli might not show these perceptual dimensions, or may indicate reliance on more than these timbral features if they were masked by the availability of musical structure information in the current stimuli. Additionally, the discrete sorting groups could invite top-down strategies based on retrieving explicit genre information from memory, and open subjective

[3] see `http://developer.echonest.com/docs/v4/_static/AnalyzeDocumentation.pdf`

experience responses will be taken in future studies to establish whether such information is cued by the clips. Nevertheless, the task is known to yield useful similarity data in a shorter and more easily administered experiment than would be possible with the more conventional pairwise similarity rating paradigm [25].

Scheirer and colleagues proposed that listeners may differ in the weight they give to a common set of perceived sound features when judging surface musical sound, or that different listeners may choose different features altogether [13]. Although they lacked enough data to explore these hypotheses, they were able to conclude that individual (participant) models explained complexity rating data better than a common model. Therefore, whilst we found some evidence of common feature-based perceptual dimensions, it is possible that further study with this paradigm will uncover individual strategy differences for this task. The INDSCAL variant of MDS may be helpful in exploring this hypothesis. The reverse is also possible, given that we used far shorter stimuli (800ms versus Scheirer et al.'s 5000) and may have measured a more constrained phenomenon. Individual differences are nonetheless plausible, as task-based measures of timbral perception can be improved by training [26, 27]. Indeed, because timbral perception does not require formalised musical knowledge, individuals could be expected to vary in the information they can access for this task purely on the basis of what they have previously listened to, and to what extent. We will look at three other datasets—including shorter, 400ms clips—and explore other features, for example those provided by Peeters and colleagues' recently published toolbox [28], as well as standard MFCC coefficients and spectral centroid-based measures.

References

1. Aucouturier, J., Pachet, F.: Improving timbre similarity: How high is the sky? In: Journal of Negative Results in Speech and Audio Sciences, vol. 1, pp. 1–13 (2004)
2. Pachet, F., Roy, P.: Exploring billions of audio features. In: Proceedings of CBMI 07, Eurasip, ed., pp. 227–235, Bordeaux, France (2007)
3. Frith, C., Horne, H.: Art into Pop. Methuen Young Books, London (1987)
4. Logan, B.: Mel frequency cepstral coefficients for music modeling. In: International Symposium on Music Information Retrieval, vol. 28, pp. 5–11 (2000)
5. Terasawa, H., Slaney, M., Berger, J.: A statistical model of timbre perception. In: ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition (SAPA2006), pp. 18–23, Pittsburgh (2006)
6. Barthet, M., Depalle, P., Kronland-Martinet, R., Ystad, S.: Analysis-by-synthesis of timbre, timing, and dynamics in expressive clarinet performance. In: Music Perception, vol. 28, pp. 265–278 (2011)
7. McAdams, S., Giordano, B.L.: The perception of musical timbre. In: The Oxford Handbook of Music Psychology, S. Hallam, I. Cross, M. Thaut, eds., pp. 72–80, Oxford University Press (2009)
8. Jehan, T.: Creating music by listening. Ph.D. thesis, Massachusetts Institute of Technology (2005)
9. Gjerdingen, R.O., Perrott, D.: Scanning the dial: The rapid recognition of music genres. In: Journal of New Music Research, vol. 37, pp. 93–100 (2008)

10. Krumhansl, C.L.: Plink: "Thin slices" of music. In: Music Perception: An Interdisciplinary Journal, vol. 27, pp. 337–354 (2010)
11. Iverson, P., Krumhansl, C.L.: Isolating the dynamic attributes of musical timbre. In: The Journal of the Acoustical Society of America, vol. 94, pp. 2595–2606 (1993)
12. Filipic, S., Tillmann, B., Bigand, E.: Judging familiarity and emotion from very brief musical excerpts. In: Psychonomic Bulletin & Review, vol. 17, pp. 335–341 (2010)
13. Scheirer, E.D., Watson, R.B., Vercoe, B.L.: On the perceived complexity of short musical segments. In: Proceedings of the 2000 International Conference on Music Perception and Cognition, Citeseer (2000)
14. Karadogan, C.: A Comparison of Kanun Recording Techniques as They Relate to Turkish Makam Music Perception. In: Proceedings of the 130th Audio Engineering Society Convention, Audio Engineering Society (2011)
15. Marui, A., Martens, W.L.: Timbre of nonlinear distortion effects: Perceptual attributes beyond sharpness. In: Proceedings of the Conference on Interdisciplinary Musicology (2005)
16. Wedin, L., Goude, G.: Dimension analysis of the perception of instrumental timbre. In: Scandinavian Journal of Psychology, vol. 13, pp. 228–240 (1972)
17. Grey, J.M.: Timbre discrimination in musical patterns. In: The Journal of the Acoustical Society of America, vol. 64, pp. 467–478 (1978)
18. Samson, S., Zatorre, R.J., Ramsay, J.O.: Deficits of musical timbre perception after unilateral temporal-lobe lesion revealed with multidimensional scaling. In: Brain, vol. 125, pp. 511–522 (2002)
19. Rentfrow, P.J., Gosling, S.D.: The do re mi's of everyday life: The structure and personality correlates of music preferences. In: Journal of Personality and Social Psychology, vol. 84, pp. 1236–1256 (2003)
20. Rentfrow, P.J., Gosling, S.D.: Message in a Ballad. In: Psychological Science, vol. 17, pp. 236–242 (2006)
21. Kruskal, J.: Nonmetric multidimensional scaling: A numerical method. In: Psychometrika, vol. 29, pp. 115–129 (1964)
22. Hastie, T., Tibshirani, R., Friedman, J.: Random Forests. In: The Elements of Statistical Learning, Springer Series in Statistics, pp. 1–18, Springer New York (2009)
23. Hothorn, T., Hornik, K., Zeileis, A.: Model-based recursive partitioning. In: Journal of Computational and Graphical Statistics, vol. 17, pp. 492–514 (2008)
24. Strobl, C., Boulesteiz, A., Kneib, T., Augustin, T., Zeileis, A.: Conditional variable importance for Random Forests. In: Bioinformatics, vol. 9, pp. 307–327 (2008)
25. Müllensiefen, D., Gingras, B., Stewart, L., Musil, J.J.: Goldsmiths Musical Sophistication Index (Gold-MSI) v0.9: Technical Report and Documentation Revision 0.2. Tech. rep., Goldsmiths, University of London, London (2012), URL `http://www.gold.ac.uk/music-mind-brain/gold-msi`
26. Shahin, A.J., Roberts, L.E., Chau, W., Trainor, L.J., Miller, L.M.: Music training leads to the development of timbre-specific gamma band activity. In: NeuroImage, vol. 41, pp. 113–122 (2008)
27. Gfeller, K., Witt, S., Adamek, M., Mehr, M., Rogers, J., Stordahl, J., Ringgenberg, S.: Effects of training on timbre recognition and appraisal by postlingually deafened cochlear implant recipients. In: Journal of the American Academy of Audiology, vol. 13, pp. 132–145 (2002)
28. Peeters, G., Giordano, B.L., Susini, P., Misdariis, N., McAdams, S.: The Timbre Toolbox: Extracting audio descriptors from musical signals. In: Journal of the Acoustical Society of America, vol. 130, pp. 2902–2915 (2011)

Towards Computational Auditory Scene Analysis: Melody Extraction from Polyphonic Music

Karin Dressler

Fraunhofer Institute for Digital Media Technology IDMT, Ilmenau, Germany
kadressler@gmail.com

Abstract. This paper describes an efficient method for the identification of the melody voice from the frame-wise updated magnitude and frequency values of tone objects. Most state of the art algorithms employ a probabilistic framework to find the best succession of melody tones. Often such methods fail, if there are several musical voices with a comparable strength in the audio mixture. In this paper, we present a computational method for auditory stream segregation that processes a variable number of simultaneous voices. Although no statistical model is implemented, probabilistic relationships that can be observed in melody tone sequences are exploited. The method is a further development of an algorithm which was successfully evaluated as part of a melody extraction system. While the current version does not improve the overall accuracy for some melody extraction data sets, it shows a superior performance for audio examples which have been assembled to show the effects of auditory streaming in human perception.

Keywords: computational auditory scene analysis, auditory stream segregation, melody extraction

1 Introduction

Melody is defined as a linear succession of musical tones which is perceived as a single entity. The melody is often the predominant voice in the sound mixture, this means it stands out from the background accompaniment. There are several features that increase the salience of the melody tone, for example loudness, frequency variation, timbre, and note onset rate. State of the art melody extraction algorithms mainly exploit two characteristics to identify the melody voice: 1) the predominance of the melody voice in terms of loudness and 2) the smoothness of the melody pitch contour.

At present two main algorithm types for the identification of the melody voice can be distinguished: on the one hand, probabilistic frameworks are used to find the optimal succession of tones. They combine pitch salience values and smoothness constraints in a cost function that is evaluated by optimal path finding methods like the hidden Markov Model (HMM) or dynamic programming (DP)

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

methods. On the other hand, there are rule based approaches that trace multiple F0 contours over time using criteria like magnitude and pitch proximity in order to link salient pitch candidates of adjacent analysis frames. Subsequently, a melody line is formed from these tone-like pitch trajectories, using rules that take the necessary precautions to assure a smooth melody contour. Of course such a division is rather artificial. It is easy to imagine a system that uses tone trajectories as input for a probabilistic framework, and vice versa a statistical approach can be used to model tones. In fact, Ryynänen and Klapuri have implemented a method for the automatic detection of singing melodies in polyphonic music, where they derive a HMM for note events from fundamental frequencies, their saliences and an accent signal [1].

Most state of the art approaches use probabilistic frameworks that accomplish the tone trajectory forming and the identification of the melody voice simultaneously [2–4]. The application of a statistical model provides an out of the box solution that evaluates different features of the melody voice, as long as they can be expressed mathematically in a cost function or a maximum likelihood function.

Rao and Rao advocate dynamic programming over variants of partial and tone tracking, but also acknowledge the drawback of current statistical approaches [3]: While for rule-based methods alternative melody lines can be recovered quite easily, there is no effective way to retrieve alternative paths using the prevailing DP approach (i.e. the Viterbi algorithm), because the mathematical optimization of the method depends on the elimination of concurrent paths. Hence, it is not easy to state whether the most likely choice stands out from all other choices.

If there is a second voice with a comparable strength in the audio mixture, the identification of the predominant voice becomes a challenging problem. Of course, this assertion is also true for rule-based methods. Unfortunately, it is not unusual to find a strong second voice in real-world music, as a booming bass line is almost mandatory in many music genres. Masataka Goto describes a system for the automatic detection of the melody and bass line for real-world music in [5]. Using realistic assumptions about contemporary music, the problem of the concurrent melody and bass line is addressed by intentionally limiting the frequency range for both voices using band pass filters. Rao and Rao present an approach towards the solution of this problem in [3], giving an example for DP with dual fundamental frequency tracking. The system continuously tracks an ordered pair of two pitches, but it cannot ensure that the two contours will remain faithful to their respective sound sources.

Another problem to be addressed is the identification of non-voiced portions, i.e. frames where no melody tones occur. The simultaneous identification of the optimal path together with the identification of melody frames is not easy to accomplish within one statistical model, so often the voicing detection is performed by a separate processing step. Nonetheless, optimal path finding algorithms may be confused by rests in the tone sequence, especially because the usual transition probabilities do not apply in between melodic phrases.

An important characteristic of the human auditory system is the influence of note onset rate on the stream segregation. Tone sequences that are a quick succession of large intervals actually fail to form a recognizable melody, since the auditory system cannot integrate the individual tones into one auditory stream [6, chapter 2]. The integration or segregation of such a tone sequence depends markedly on the duration of the tones, so a voice processing algorithm should take into account such temporal aspects, too.

In this paper, we present an algorithm for the identification of the predominant voice in music that addresses some of the problems mentioned above. An auditory streaming model is implemented, which takes the frame-wise frequency and magnitude information of tones as input. With this information, so-called voice objects are established, which in turn capture salient tones close to their preferred frequency range. Although no statistical model is implemented, probabilistic relationships that can be observed in melody tone sequences are exploited. The presented method is a further development of an algorithm presented in [7]. The main technical difference over the baseline method is the renunciation of the mediated tone search using streaming agents. In the updated version, the voice object itself actively seeks the next voice tone. This is a big advantage, because – supplemental to increased algorithm performance – additional (voice dependent) search criteria can be integrated, like for example timbral features.

2 Statistical Properties of Melodies

By voices musicians mean a single line of sound, more or less continuous, that maintains a separate identity in a sound field or musical texture. The melody has certain characteristics that establish it as the predominant voice in the musical piece. Of course, a musical voice is not a succession of random notes – tones belonging to the same voice usually have a similar timbre, intervals between notes have a certain probability, there are rules regarding harmony and scale, and onset times of notes can be related to a rhythmical pattern.

Unfortunately the retrieval of high level musical features from polyphonic music is a challenging task in itself. Even for the most prominent voice (i.e. the melody), it is difficult to identify note onsets or to assign a note name to a tone with a varying frequency.

However, a melodic succession of tones has statistical properties that can be more easily exploited. Huron states that pitch proximity is the best generalization about melodies [8, chapter 5]. This statement is well supported by the interval statistics[1], as melodies consist mostly of tone sequences that are typically close to one another in pitch (see figure 1). Indeed, the unison is the most frequent interval by a great margin, followed by the whole tone interval.

[1] The Fraunhofer Institute in Ilmenau has gathered a collection of 6000 MIDI songs containing multiple genres, ranging from classical to contemporary charts music. Nearly one million notes were analyzed to compile a statistic of interval occurrences and the average note durations in melody tone sequences.

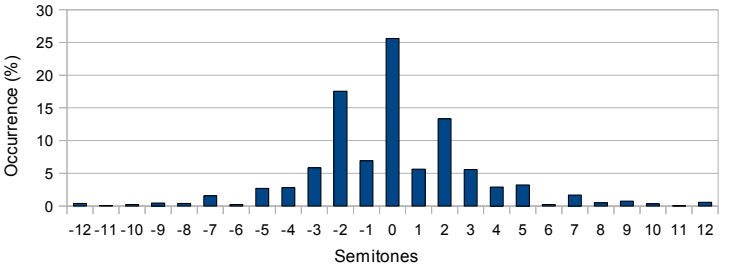


Fig. 1. Histogram of Note Intervals in Melodies

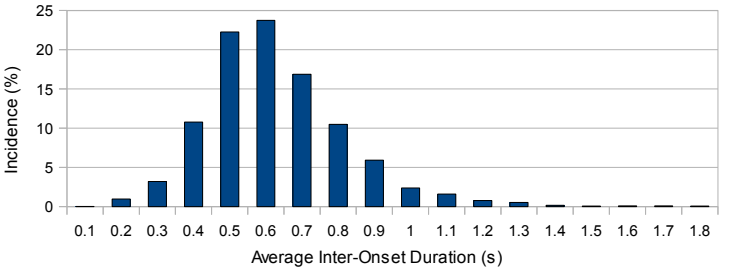


Fig. 2. Histogram of the Average Note Duration in Melodies

Other essential cues that help to distinguish musical voices are the central pitch tendency and regression to the mean [8, chapter 5]: the most frequently occurring pitches lie near the center of the melody's range. A necessary consequence of this tendency is the fact that after a melodic leap (an interval of more than three semitones) away from the center of the tone distribution, the following interval will change direction with a high probability. Regression to the mean is the most general explanation for this post-leap reversal.

The duration of melody tones lies normally in the range of 150 to 900 ms (see figure 2). Notes at faster rates occur, but they usually do not contribute to the perception of melody [9, chapter 5]. If a familiar tune is played at a rate faster than approximately 50 ms per note, the piece will not be recognizable, although the global melodic contour can be perceived. Yet, a very slow playback (i.e. durations of more than one second) is possible.

The dynamic range, which denotes the ratio between the largest and smallest occurring magnitudes in a tone sequence, is another important cue. Usually, tones that belong to the same voice have more or less the same sound level. It should be noted, however, that especially the human singing voice has a rather high dynamic range with ratios of more than 20 dB between the loudest tones and the softest ones.

The process that is required by the human auditory system as it analyzes mixtures of simultaneous and sequential sound entities has been coined auditory

scene analysis [6]. All of the aforementioned statistical properties of melodies in fact enable the sequential grouping of sounds by the human auditory system. These "primitive" grouping principles are not only valid for music, but also for speech, environmental sounds, and even for noise.

Still, the ability of humans to distinguish concurrently sounding voices is limited. Huron investigates the ability of musically trained listeners to continuously report the perceived number of voices in a polyphonic musical performance in [10]. While Huron questions the musical significance of his experiment, because it does not evoke a natural listening situation, one important take away is that there is a marked worsening of the human performance, when a three-voice texture is augmented to four voices. If errors occur, the number of voices is underestimated in 92 percent of the cases. Another finding of the experiment is the fact that inner voices are more difficult to detect. The reaction time for the identification of an inner voice is twice as long, and often they are not detected at all.

3 Method

The formation of voices is controlled by the frame-wise updated magnitude and frequency of tone objects, which have a fundamental frequency in the range between 55 and 1319 Hz. The time advance between two successive analysis frames denotes 5.8 ms. Tone objects can be seen as pitch trajectories derived from salient pitches in so-called pitch spectrograms which may be computed with diverse pitch determination algorithms (PDA) like for example [11, 12]. Most PDA do not only compute pitch frequencies, but also offer an estimate for the corresponding pitch strengths, which is used as tone magnitude.

3.1 Overview

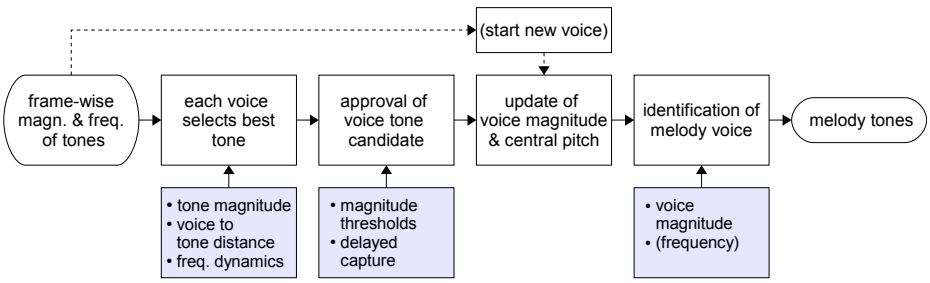


Fig. 3. Algorithm Overview

Figure 3 shows the processing steps performed in each analysis frame (i.e. every 5.8 ms). The input to the algorithm are the magnitude and frequency of the tone objects. The starting point of a new voice object is a salient tone which has not been added to an existing voice. In each analysis frame, every voice independently selects one tone, preferring strong tones that are close to its central pitch. If the selected tone passes all magnitude thresholds, it is added to the voice (after a certain delay period). The magnitude and central pitch of the voice are updated, whenever it has an added voice tone: the voice assembles a magnitude corresponding to the magnitude of the captured tone, and at the same time the voice's central pitch gradually moves towards the pitch of the added tone. Finally, the melody voice is chosen from the set of voices. The main criterion for the selection is the magnitude of the voice. Only tone objects of the melody voice qualify as melody tones.

3.2 Start Conditions

The first question to ask is at which point a new voice should be started[2]. The conditions for starting a new voice object are as follows:

- A voice is started from a tone which was not included in an existing voice.
- The tone reached at least once the maximum magnitude among all other tones.
- The magnitude of the tone has passed at least once the global magnitude threshold.
- There is no voice which could capture the tone, or the duration of the tone is greater than 200 ms, or the tone was finished.

3.3 Selection of Voice Tone Candidates

In each analysis frame, the voice object searches for a strong tone in the frequency range of ±1300 cent around its current central pitch. The best choice, at the one hand, ensures the smoothness of the voice tone sequence, at the other hand, embraces tones with a strong magnitude. In contrast to most existing approaches using optimal path finding methods, the smoothness of the melody line is evaluated in terms of central pitch, and not with respect to the last added tone. This strategy might not give the best results in every situation, but it reinforces the importance of the central pitch, and allows an easier recovery after an erroneous addition of a tone.

The Rating of Voice Tone Candidates: Each voice independently chooses only one tone – the object with the maximum rating A_{rating}:

$$A_{\text{rating}} = C \cdot D \cdot A_{\text{tone}} \cdot g_1(\Delta c) \tag{1}$$

[2] The conditions given here are crafted for the purpose of melody extraction, which aims at the identification of the predominant melody line. If voices besides the predominant one shall be extracted, it is advisable to define more inclusive conditions.

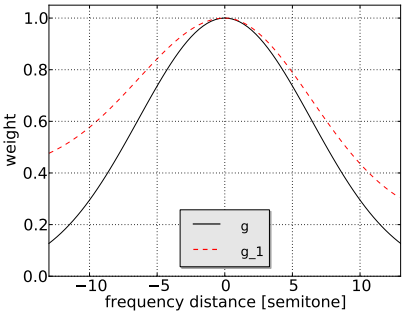


Fig. 4. Weighting Functions

The rating is calculated from the following four criteria:

- *Magnitude:* The tone magnitude A_{tone} is a good indicator for the perceptual importance of a tone.
- *Frequency distance weight:* The voice should preferably select a tone that is close to its central pitch. That is why the magnitude of the tone is weighted by a function that takes into account the frequency distance Δc between the tone's pitch c_{tone} and the central pitch of the voice $\bar{c}_{\text{voice}}$:

$$g_1(\Delta c) = r + (1 - r) \cdot g(\Delta c) \qquad \text{with } \Delta c = c_{\text{tone}} - \bar{c}_{\text{voice}}. \tag{2}$$

  The parameter $r = 0.4$ if Δc is negative, otherwise $r = 0.2$, and g is the function

$$g(\Delta c) = e^{-0.5\frac{(\Delta c)^2}{640^2}}. \tag{3}$$

  Figure 4 shows that the resulting weighting function g_1 is asymmetric – the weighting is biased towards tones from the lower frequency range. There are two reasons for this asymmetry. First, overtone errors cannot be avoided entirely, so in doubt the lower pitch is probably the true fundamental frequency. Second, tones in the lower frequency range of an instrument or the human voice are often softer, so the weighting compensates this difference.
- *Comparison with average magnitude:* The magnitude of the selected tone candidate should be in the order of the previously added magnitudes. For the comparison we use the maximum tone magnitude $\hat{A}_{\text{tone}}$ and the long term exponential moving average (EMA) of the maximum tone magnitudes of previously added tones[3]. If $\hat{A}_{\text{tone}}$ is more than 10 dB below or above the long term average, the rating is halved. Accordingly a magnitude factor C is set to 1 or 0.5 in the final rating.
- *Frequency deviation:* Sounds with changing attributes attract attention. Human listeners particularly focus on tones with vibrato or pitch glides. If a

[3] A detailed description of the exponential moving average can be found in the appendix.

tone shows persistently more than 20 cent frequency difference in between analysis frames the rating is doubled. Accordingly, a deviation factor D is set to 1 or 2 in the final rating.

Different Voices Competing for the Same Tone Any tone object preferably belongs to only one voice. In practice, there are often ambiguous situations, where an exclusive assignment to one voice is not the optimal solution.

The priority is on voices with a larger voice magnitude: a previously added tone may still be added by another voice, if the new owner has a larger magnitude than the current owner of the tone. Having said that, any voice which has a smaller magnitude than the current owner of the tone is prohibited to add the tone. The priority on strong voices is also reflected in the selection of the tones which was described in the previous section. The aim is that weaker voices shall avoid tones that are already added to strong voices (i.e. a voice with a large magnitude $\bar{A}_{\text{voice}}$). Hence, two more rating factors are introduced to direct the attention of weak voices to other suitable tone candidates:

- *Comparison of voice magnitude:* Whenever the tone is already included in a stronger voice, the original rating A_{rating} is multiplied with the factor 0.7.
- *Comparison of voice bidding:* If two voices aim at the same tone, the rating A_{rating} is decreased by the factor 0.7 for the voice with the lower bidding, but only if it is also the weaker voice. The voice bidding is the product of voice magnitude and the distance weight given in equation 2: $A_{\text{bidding}} = \bar{A}_{\text{voice}} \cdot g_1(\Delta c)$.

As the voice magnitudes and the voice biddings of the current frame are not known prior to the tone selection process, the values of the last analysis frame are used for the comparison. As the values usually change rather slowly, they are still significant. Furthermore, this provision ensures that the output is independent of the explicit order in which voices bid for tones.

3.4 Approval of Voice Tones

Even though one voice tone candidate is selected in each analysis frame, it is not clear whether the particular tone belongs to the voice or not, as melodies also contain rests. Two different techniques are employed to perform the voicing detection, namely the use of adaptive magnitude thresholds and the delayed capture of tones.

Short Term Magnitude Threshold The short term magnitude threshold is estimated for each voice individually. It secures that shortly after a tone is finished no weak tone is added to the voice prematurely. Hence, it is especially useful to bridge small gaps between tones of a voice. The short term threshold is adaptive and decays with a half-life time of 150 ms. Whenever the current voice

tone has a magnitude which is larger than the current threshold reference value $T_{150\text{ms}}$, it is updated to the new maximum:

$$T_{150\text{ms}} \leftarrow \begin{cases} A_{\text{tone}}, & \text{if} \quad A_{\text{tone}} > T_{150\text{ms}}; \\ \alpha_{150\text{ms}} \cdot T_{150\text{ms}}, & \text{otherwise.} \end{cases} \tag{4}$$

The parameter $\alpha_{150\text{ms}}$ controls the decay of the magnitude threshold. The calculation of its value is described in equation 13. The tone passes the threshold if it is no more than 6 dB below $T_{150\text{ms}}$.

Long Term Magnitude Threshold The long term magnitude threshold $T_{5\text{s}}$ is basically the same as the short term threshold, with the distinction that it decays with a half-life period of 5 seconds. In order to pass the threshold, the tone's magnitude should not be more than 20 dB below $T_{5\text{s}}$.

Long Term EMA Magnitude Threshold A high dynamic range of 20 dB within a tone sequence is not exceptional – a prominent example is the human singing voice. However, if a relatively high dynamic range is allowed, many tones from the accompaniment will pass the magnitude threshold, too. This is especially true for instrumental music, which often contains several simultaneous voices with a comparable strength. Besides the long term threshold, which is based on the maximum magnitude, another threshold is introduced which is computed as the exponential moving average of the previously added voice tone magnitudes. This threshold is updated whenever the voice has an approved voice tone, provided that the tone's duration is between 50 and 500 ms.

$$T_{\text{EMA\_5s}} \leftarrow \alpha_{5\text{s}} \cdot T_{\text{EMA\_5s}} + (1 - \alpha_{5\text{s}}) \cdot \hat{A}_{\text{tone}} \tag{5}$$

The EMA is estimated with the current peak magnitude $\hat{A}_{\text{tone}}$, which denotes the biggest magnitude the tone has reached so far. At the start of the voice the magnitude threshold is set to one third of the maximum magnitude of the first added voice tone. As the threshold reflects the dynamic range of previous voice tone magnitudes, the actual threshold value can be defined more strictly. In order to pass the threshold, the tone's magnitude should not be more than 10 dB below $T_{\text{EMA\_5s}}$.

Delayed Capture of Tone The approval of a new voice tone is often delayed to allow some time for the start of a more suitable tone. The delay time depends on the distance between the candidate voice tone and the preferred frequency range of the voice (see section 3.5). All tones within the preferred frequency interval are added immediately, provided that they pass the magnitude thresholds. All other tones face a delay that depends on their magnitude and the frequency distance between tone and the preferred frequency range.

In order to estimate the delay, a short term pitch $\bar{c}_{\text{st}}$ is defined for each voice object. (The computation of $\bar{c}_{\text{st}}$ is described in section 3.5.) The tone may

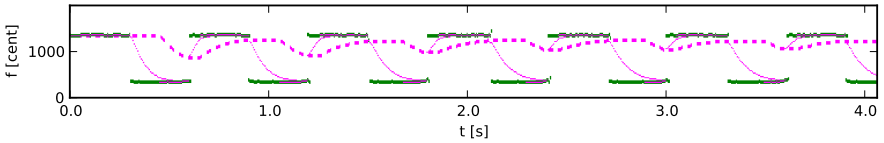


Fig. 5. Alternating tones: dashed line - central pitch of the voice $\bar{c}_{\text{voice}}$, thin line - short term pitch of the voice $\bar{c}_{\text{st}}$.

only be added after $\bar{c}_{\text{st}}$ has approximately reached the frequency of the voice tone candidate (i.e. less than 100 cent distance). Figure 5 illustrates the delayed capture of alternating tones.

3.5 Update of Voice Parameters

Contrary to the baseline method presented previously in [7], the intermediate step of streaming agents is omitted in this implementation. Consequently, voice objects do not derive their magnitude and central pitch from the assigned streaming agent. In the presented approach, the voice parameters are calculated directly based on the added tones.

Magnitude Update The voice magnitude $\bar{A}_{\text{voice}}$ is updated whenever the voice has an approved voice tone. The magnitude depends on the tone's rating magnitude A_{rating} as given in equation 1. The use of the rating magnitude ensures that a voice profits more from tones which are close to its current central pitch. In order to update the magnitude values, we use the exponential moving average (EMA).

$$\bar{A}_{\text{voice}} \leftarrow \alpha_{500\text{ms}} \cdot \bar{A}_{\text{voice}} + (1 - \alpha_{500\text{ms}}) \cdot A_{\text{rating}}. \tag{6}$$

The parameter $\alpha_{500\text{ms}}$ is a smoothing factor which corresponds to a half-life period of 500 ms. The EMA calculation is initialized with a fraction of the peak magnitude of the first tone: $\bar{A}_{\text{voice}} = 0.2 \cdot \hat{A}_{\text{tone}}$.

Central Pitch The central pitch of the voice $\bar{c}_{\text{voice}}$ is an important parameter, as it defines the preferred frequency range for the selection of tones. It is established over time according to the pitches of approved voice tones. While the adaptation could be implemented as EMA of previous frequencies, it is beneficial if the adaptation speed also depends on the tone's magnitude. This means the central pitch moves faster towards strong tones. That is the reason why at first a weight $\bar{A}_{\text{w}}$ is defined, which allows to evaluate the current rating of a tone in relation to the EMA of previous ratings:

$$\bar{A}_{\text{w}} \leftarrow \left(\bar{A}_{\text{w}} - A_{\text{rating}}\right) \alpha_{500\text{ms}} + A_{\text{rating}} \tag{7}$$

The EMA is initialized with $\bar{A}_{\mathrm{w}} = 0.2 \cdot \hat{A}_{\mathrm{tone}}$ at the beginning of the voice. With the help of the weight $\bar{A}_{\mathrm{w}}$ we can finally update the central pitch:

$$\bar{c}_{\mathrm{voice}} \leftarrow \frac{\bar{A}_{\mathrm{w}} \bar{c}_{\mathrm{voice}} + (1 - \alpha_{500\mathrm{ms}}) \cdot A_{\mathrm{rating}} \cdot c_{\mathrm{tone}}}{\bar{A}_{\mathrm{w}} + (1 - \alpha_{500\mathrm{ms}}) \cdot A_{\mathrm{rating}}}. \tag{8}$$

The parameter $\alpha_{500\mathrm{ms}}$ is a smoothing factor, which corresponds to a half-life period of 500 ms[4]. The parameter A_{rating} refers to the rating magnitude of the voice tone as given in equation 1, while c_{tone} is the pitch of the voice tone. The initial value for the iterative calculation is the frequency of the first added voice tone: $\bar{c}_{\mathrm{voice}} = c_{\mathrm{tone}}$. As $\bar{A}_{\mathrm{w}}$ is close to zero at the start of the voice, the central pitch changes more rapidly after the start of the voice (see also figure 5). This is, however, a deliberate decision, as the "true" central pitch has to be established over a longer time period.

Sometimes the frequency of a tone sequence does not prevail close to a central pitch, but moves upwards or downwards in one direction. As the central pitch adapts quite slowly, the update might not be fast enough to capture the succession of tones, and soon the tones fall outside the maximum search range of the voice. To avoid this, there is an immediate update of the central pitch, if $|\bar{c}_{\mathrm{voice}} - c_{\mathrm{tone}}| > 900$. In this case the central pitch is set to the maximum distance of 900 cent.

Frequency Range Intervals which are greater than an octave are rarely found in melody tone sequences. Consequently the search range for voice tones is limited to the range of ±1300 cent around the central pitch of a voice.

Moreover, a preferred frequency range R_{pref} is defined, which is given by the frequency range between the last added voice tone frequency and the central pitch of the voice.

Short Term Pitch The short term pitch $\bar{c}_{\mathrm{st}}$ seeks to emulate the time that is needed to focus attention to a tone that is outside the preferred frequency range of the voice. It is updated whenever the voice tries to capture a new voice tone, so it is updated even without an approved voice tone.

The short term pitch $\bar{c}_{\mathrm{st}}$ can immediately be set to any frequency within the preferred frequency range R_{pref}. So if the distance to a voice tone candidate can be decreased by changing the short term pitch to a frequency within R_{pref}, $\bar{c}_{\mathrm{st}}$ is set to that value. Apart from that, the short term pitch is updated very much like the central pitch of the voice – namely by using a weighted EMA. At first, a weight $A_{\mathrm{w\_st}}$ is defined, which allows to compare the tone's current rating A_{rating} with the magnitude of previously added voice tones. For this purpose we determine $A_{\mathrm{w\_st}}$ as the average of the long term EMA magnitude threshold

[4] Since the weight $\bar{A}_{\mathrm{w}}$ depends on many factors, the parameter α does not exactly set any half-life period for the central pitch update. Yet the corresponding time span gives a reference point for the approximate adaptation speed.

$T_{\mathrm{EMA\_5s}}$ and the short term magnitude threshold $T_{150\mathrm{ms}}$:

$$A_{\mathrm{w\_st}} = 0.5 \cdot (T_{\mathrm{EMA\_5s}} + T_{150\mathrm{ms}}). \tag{9}$$

Finally, the short term pitch is updated using a weighted EMA:

$$\bar{c}_{\mathrm{st}} \leftarrow \frac{A_{\mathrm{w\_st}}\bar{c}_{\mathrm{st}} + (1 - \alpha_{30\mathrm{ms}}) \cdot A_{\mathrm{rating}} \cdot c_{\mathrm{tone}}}{A_{\mathrm{w\_st}} + (1 - \alpha_{30\mathrm{ms}}) \cdot A_{\mathrm{rating}}}. \tag{10}$$

The parameter $\alpha_{30\mathrm{ms}}$ is again the smoothing factor. Figure 5 shows how $\bar{c}_{\mathrm{st}}$ is used to capture tones: only if the thin line reaches the voice tone candidate (i.e. less than 100 cent distance), the tone may be added to the voice.

3.6 The Identification of the Melody Voice

The most promising feature to distinguish melody tones from all other sounds is the magnitude. The magnitude of the tones is of course reflected by the voice magnitude. Hence, the voice with the highest magnitude is in general selected as the melody voice. It may happen that two or more voices have about the same magnitude and thus no clear decision can be taken. In this case, the voices are weighted according to their frequency: voices in very low frequency regions receive a lower weight. The magnitude thresholds are defined for each voice individually. As they depend solely on the past tones of the voice, they cannot take effect on all soft tones. Therefore, it is recommended that a global magnitude threshold is estimated from the identified melody tones. Subsequently, the melody tones should be compared to the global threshold.

4 Results

4.1 Audio Melody Extraction

The presented method for the identification of musical voices has been implemented as part of a melody extraction algorithm which was evaluated using the melody extraction training data sets of ISMIR 2004 and MIREX 2005. Algorithm parameters regarding the width and the shape of the weighting functions as well as the timing constants of the adaptive thresholds have been adjusted using the same data sets. The previous algorithm version, which has been described in [7], is used as a benchmark[5]. The comparison with the previous algorithm version (kd2009) shows that the overall accuracy is not improved by the new method (kd2011)(see table 1). However, the results of the previous algorithm should not be seen as a baseline, as it still can be considered as a state of the art algorithm. Table 2 shows that its overall melody extraction accuracy is close to the best algorithm of the most recent MIREX audio melody extraction task, which was submitted by Salamon and Gómez [14].

[5] The melody extraction algorithm using the previous voice detection method was evaluated at the Music Information Retrieval Evaluation eXchange (MIREX) [13].

Table 1. Comparison of Melody Extraction Results for the Training Datasets

Dataset	Algorithm	Overall Accuracy (%)
ADC 2004	kd 2009	**89.2**
	kd 2011	87.5
MIREX train '05	kd 2009	73.9
	kd 2011	**74.3**

Table 2. Melody Extraction Results of MIREX 2009 (4 best submissions and the best submission of MIREX 2011)

Algorithm	Voicing Recall (%)	Voicing False Alarm (%)	Raw Pitch (%)	Overall Accuracy (%)	Runtime (min)
kd	90.9	41.0	80.6	73.4	24
dr1	92.4	51.7	74.4	66.9	23040
dr2	87.7	41.2	72.1	66.2	524
rr	91.3	51.1	72.2	65.2	26
sg (2011)	-	-	-	**75**	-

One problem of the evaluation is that a melody extraction system is only interested in the predominant voice. The previous algorithm, as well as optimal path finding algorithms, already gives satisfactory results as long as the melody voice is indeed predominant. If the audio signal contains concurrent voices of comparable strength, it is important that all strong voices are retrieved, so that the final decision can be based on a more complete picture of the audio input. A qualitative comparison of the algorithm outputs allows a more meaningful evaluation than numbers alone.

4.2 Qualitative Analysis

A qualitative analysis of the results confirms that the new method has indeed some advantages over the baseline method:

- The minimum distance between two voices is decreased. (see figures 6 - 8)
- The detection of weak voices is improved. (see figures 6 - 8)
- The behavior of the algorithm is closer to human perception, when artificial audio examples for auditory stream segregation are used as input. (see for example figures 5 and 9)
- The implementation of the new method is more straight forward, because the intermediate processing of so-called streaming agents, was omitted (see reference [7]).
- The proposed method allows a simpler inclusion of timbral features, as the voice tone candidates are directly selected by the voice and not by streaming agents.
- The computation time for the voice processing scales with the complexity of the audio input.

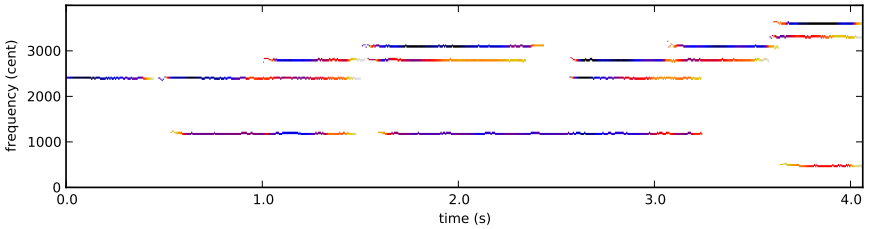


Fig. 6. Midi3.wav from the ADC 2004 data set is an example for instrumental music with several concurrent voices. The figure shows the identified tone objects, which constitute the input to the voice processing algorithm. The melody voice is in the high frequency range. The melody voice is the predominant voice, but the bass voice has a comparable strength.

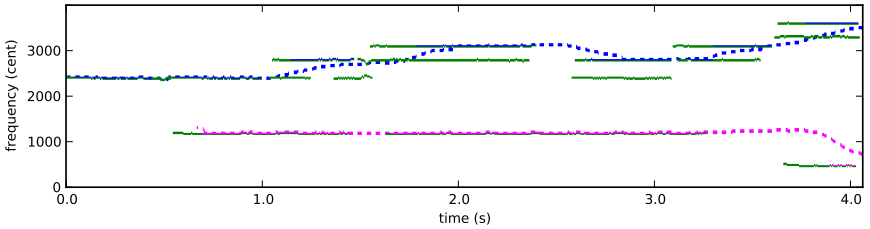


Fig. 7. Baseline method: When the bass voice starts, a second voice object is created. Two voices are recognized – the melody voice (blue) and the bass voice (magenta).

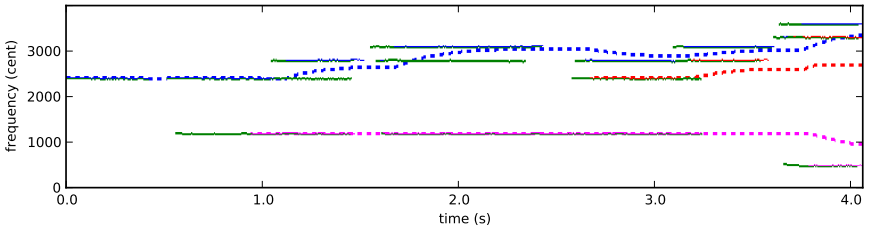


Fig. 8. Proposed method: Three voices are recognized – the melody voice (blue), the bass voice (magenta), and an inner voice (red).

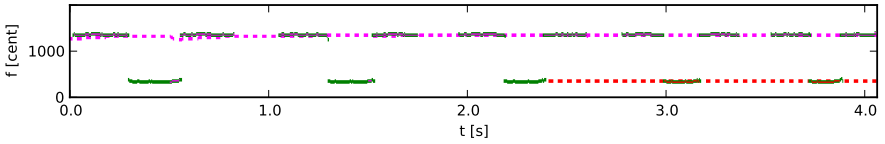


Fig. 9. Example of alternating tones: The duration of the tones is decreased over time. Soon the alternating tones cannot be captured by the first voice and a second voice is started.

5 Summary and Conclusion

In this paper, we presented an efficient approach to auditory stream segregation for melody extraction algorithms. The proposed method allows a reliable identification of a variable number of simultaneous voices in different kinds of polyphonic music. The qualitative comparison with a previous implementation shows that the proposed method improves the detection of musical voices. Furthermore, the new approach offers more possibilities to add voice dependent features for the tone selection in future implementations. Taking into account not only the magnitude and the occurring frequency intervals, but also the duration of tones, the presented algorithm is another step towards auditory stream segregation as performed by the human auditory system.

6 Appendix

6.1 Exponential Moving Average

A simple moving average is the mean of the previous N data points. An exponential moving average (EMA) applies weighting factors to all previous data points which decrease exponentially, giving more importance to recent observations while still not discarding older observations entirely. The smoothing factor α determines the impact of past events on the actual EMA. It is a number between 0 and 1. A lower smoothing factor discards older results faster.

The computation of the EMA can be expressed by the following formula

$$\bar{y}_l = (1-\alpha)\sum_{i=0}^{l-1}\alpha^i y_{l-i}, \tag{11}$$

where l designates the current time period (i.e. current analysis frame), y_l is the current observation, and $\bar{y}_l$ the resulting EMA value.

However, the application of equation 11 is inconvenient, because all previous data samples have to be weighted and summed in order to compute the EMA. The same result can be achieved using the following recursive formula for time periods $l > 0$:

$$\bar{y}_l = \alpha \cdot \bar{y}_{l-1} + (1-\alpha)\cdot y_l. \tag{12}$$

Equation 12 shows that the EMA can be calculated very efficiently from only two numbers: the current observation data y_l and the preceding EMA value $\bar{y}_{l-1}$. Thus, a big advantage of this method is that no previous data has to be stored in memory (besides the last EMA value).

In order to make the first recursive computation possible, the EMA value has to be initialized. This may happen in a number of different ways. Most commonly $\bar{y}_0$ is initialized with the value of the first observation. The problem of this technique is that the first observation gains a huge impact on later EMA results. As another option, the first EMA value can be set to 0. In this case the observations have comparable weights, but the calculated EMA values do not

represent an average of the observations. Rather, the EMA value starts close to zero and then approaches the average slowly – just like a sampled capacitor charging curve.

For the actual implementation it is important to figure out optimal values for the smoothing factor α. A more intuitive measure than the smoothing factor is the so-called half-life period. It denotes the time span needed to decrease the initial impact of an observation by a factor of two. Taking into account the desired half-life t_h and the time period between two EMA calculations Δt, the corresponding smoothing factor is calculated as follows:

$$\alpha = 0.5^{\frac{\Delta t}{t_h}}. \tag{13}$$

References

1. M. Ryynänen and A. Klapuri. Transcription of the singing melody in polyphonic music. In *Proc. of the 7th International Society for Music Information Retrieval Conference (ISMIR)*, Victoria, Canada, Oct. 2006.
2. C.-L. Hsu, L.-Y. Chen, J.-S. R. Jang, and H.-J. Li. Singing pitch extraction from monaural polyphonic songs by contextual audio modeling and singing harmonic enhancement. In *Proc. of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, Kobe, Japan, Oct. 2009.
3. V. Rao and P. Rao. Improving polyphonic melody extraction by dynamic programming based dual f0 tracking. In *Proc. of the 12th International Conference on Digital Audio Effects (DAFx)*, Como, Italy, Sept. 2009.
4. J.-L. Durrieu, G. Richard, B. David and C. Fvotte Source/Filter model for unsupervised main melody extraction from polyphonic audio signals. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3):564–575, Mar. 2010.
5. M. Goto. A real-time music-scene-description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals. *Speech Communication (ISCA Journal)*, 311–329, Sept. 2004.
6. A. S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*, volume 1 MIT Press paperback. MIT Press, Cambridge, Mass., Sept. 1994.
7. K. Dressler An auditory streaming approach for melody extraction from polyphonic music. In *Proc. of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, Miami, Florida, Oct. 2011.
8. D. Huron. *Sweet Anticipation: Music and the Psychology of Expectation*. The MIT Press, Cambridge, Massachusetts, 2006.
9. R. M. Warren. *Auditory perception: an new analysis and synthesis*. Cambride University Press, 1999.
10. D. Huron. Voice denumerability in polyphonic music of homogeneous timbres. *Music Perception*, 6(4):361–382, 1989.
11. K. Dressler Pitch estimation by the pair-wise evaluation of spectral peaks. In *AES 42nd Conference*, Ilmenau, Germany, July 2011.
12. D.J. Hermes Measurement of pitch by subharmonic summation *Journal of the Acoustical Society of America*, 83(1):257–264, 1988.
13. K. Dressler Audio Melody Extraction for MIREX 2009. In *5th Music Information Retrieval Evaluation eXchange (MIREX)*, 2009.
14. J. Salamon and E. Gómez Melody extraction from polyphonic music: MIREX 2011. In *7th Music Information Retrieval Evaluation eXchange (MIREX)*, 2011.

Poster session 2:

Computer Models of Music Perception and Cognition, Music Information Retrieval, Music Similarity and Recommendation, Musicology, Intelligent Music Tuition Systems

Predicting Emotion from Music Audio Features Using Neural Networks

Naresh N. Vempala and Frank A. Russo

SMART Lab, Ryerson University
nvempala@psych.ryerson.ca

Abstract. We describe our implementation of two neural networks: a static feedforward network, and an Elman network, for predicting mean valence/arousal ratings of participants for musical excerpts based on audio features. Thirteen audio features were extracted from 12 classical music excerpts (3 from each emotion quadrant). Valence/arousal ratings were collected from 45 participants for the static network, and 9 participants for the Elman network. For the Elman network, each excerpt was temporally segmented into four, sequential chunks of equal duration. Networks were trained on eight of the 12 excerpts and tested on the remaining four. The static network predicted values that closely matched mean participant ratings of valence and arousal. The Elman network did a good job of predicting the arousal trend but not the valence trend. Our study indicates that neural networks can be trained to identify statistical consistencies across audio features to predict valence/arousal values.

Keywords: valence, arousal, music, emotion, machine learning, neural networks.

1 Introduction

A common reason for engaging in music listening is that music is an effective means of conveying and evoking emotions. Although these emotions may be subjective, based in part on the listener's cultural and musical background, there are commonalities in perceived emotion across different listeners based on the characteristics of the music. Several studies have attempted to predict emotion conveyed during music listening. Some studies have explored the relationship between physiological activity experienced by a listener and perceived emotion [1-2]. Others have explored the relationship between perceived emotion and the musical/acoustic features themselves [3-4]. While acknowledging that individual differences exist in the emotion conveyed by any one piece of music, we believe that it is legitimate to consider the modal appraisal and that this appraisal may be predicted on the basis of features extracted from the music.

Various methods have been used to represent emotion perceived by listeners. One common method, described by Russell's circumplex [5], involves representing emotion using a two-dimensional space with valence on the x-axis and arousal on the y-axis. Schubert [3] used the circumplex model to identify the relationship between musical features and perceived emotion. Changes in loudness and tempo were

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

positively correlated with changes in arousal, and melodic contour was positively correlated with valence. A few studies [4, 6] have used machine-learning techniques to predict discrete emotion categories based on audio features in musical excerpts. Laurier et al. [4] extracted timbral, tonal, and rhythmic audio features from film soundtrack excerpts that were evaluated by participants, for five different emotions. Based on this data, Laurier et al. used Support Vector Machines to classify excerpts into the five discrete emotions.

As opposed to classifying a musical excerpt into a discrete emotion category, our aim was to apply machine-learning techniques towards predicting valence and arousal values on the two-dimensional emotion space based on Russell's circumplex. A nonlinear regression function that predicts valence/arousal values offers a significant contribution to existing methods relating audio-based features to perceived emotion because, there are situations when participants may be unclear on the type of emotion conveyed by the music due to the overlapping and/or ambiguous nature of some emotions. In such cases, dimensional ratings provide a more effective means of representing the emotion conveyed by the music. We used machine learning, specifically feedforward neural networks, for predicting ratings on valence and arousal dimensions.

Although neural networks have been applied extensively in domains such as object recognition, speech and text recognition, they have been relatively underutilized in music cognition and music informatics. We designed two separate networks to predict listeners' mean valence and arousal ratings associated with musical excerpts. The first network was a standard, static feedforward neural network designed to predict valence and arousal ratings for the entire excerpt. The second network was an Elman network designed to predict valence and arousal ratings for 30-second increments of the excerpt while using the previous 30 seconds as context, to understand how context might influence ratings over time.

2 Feature Extraction and Data Collection

We used 12 classical music excerpts from 12 different composers as stimuli (see Table 1). Each excerpt lasted 120 seconds. These excerpts were selected such that three excerpts represented each of the four emotion quadrants in Russell's circumplex: high arousal, positive valence (*Happy),* high arousal, negative valence (*Agitated)*, low arousal negative valence (*Sad)*, and low arousal, positive valence (*Peaceful)*. Excerpts were chosen based on previous work investigating emotional responses to music [7-8]. Using MIRtoolbox [9], we extracted 13 low- and mid-level features pertaining to dynamics, rhythm, timbre, pitch and tonality: *rms*, *lowenergy*, *eventdensity*, *tempo*, *pulseclarity*, *zerocross*, *centroid*, *spread*, *rolloff*, *brightness*, *irregularity*, *inharmonicity*, and *mode*. Values of all the features were normalized between 0 and 1. For the standard feedforward neural network, we used data from 45 participants (37 females, 2 males, 6 unknown; M_{age}=24.8, SD_{age}=8.2) with limited musical training. Participants heard each excerpt and rated two dimensions of emotion: unpleasant vs. pleasant (i.e., valence) and calm vs. excited (i.e., arousal) on a scale from 1 (least pleasant/least excited) to 9 (most pleasant/most excited). For the Elman network, we collected data from 9 participants (5 females, 4 males; M_{age}=30.4,

SD_{age}=6.8) with music training. Each of the 12 excerpts lasting 120 seconds was broken into four sequential segments of equal duration totaling 48 separate segments. For each excerpt, participants heard the four 30-second segments in sequence. After each segment, they provided their valence/arousal ratings following the same procedure as was used for the previous 45 participants. For the second, third, and fourth sequential segments of each melody, participants were told to assume that these segments were a continuation of the previous segment when providing their ratings. The same 13 features were extracted from each of the 48 segments using MIRtoolbox.

Table 1. 12 music excerpts with composers, emotion quadrants, and mean participant ratings.

Excerpt	Composer	Composition	Quadrant	Mean Arousal	Mean Valence
M1	Bartok	Sonata for 2 pianos and percussion (Assai lento)	Agitated	5.9	4.8
M2	Shostakovich	Symphony No. 8 (Adagio)	Agitated	7.1	4.1
M3	Stravinsky	Danse sacrale (Le Sacre du Printemps)	Agitated	6.6	4.1
M4	Beethoven	Symphony No. 7 (Vivace)	Happy	5.8	6.2
M5	Liszt	Les Preludes	Happy	6.4	6.0
M6	Strauss	Unter Donner und Blitz	Happy	7.0	6.8
M7	Bizet	Intermezzo (Carmen Suite)	Peaceful	2.5	6.3
M8	Dvorak	Symphony No. 9 (Largo)	Peaceful	2.4	5.7
M9	Schumann	Traumerei	Peaceful	2.9	5.2
M10	Chopin	Funeral March, Op. 72 No. 2	Sad	2.5	4.8
M11	Grieg	Aase's Death (Peer Gynt)	Sad	4.1	3.9
M12	Mozart	Requiem (Lacrimosa)	Sad	3.4	4.4

3 Methods

In this section we describe the linear and nonlinear regression methods that were used to (a) examine the relationship between audio features and valence/arousal ratings, and (b) predict valence/arousal ratings based on audio features[1].

3.1 Correlation of Audio Features with Emotion Ratings

As a first step towards understanding the pattern by which audio features might account for emotion ratings, we conducted correlational analyses between features and mean valence/arousal ratings of the 45 participants for the 12 excerpts. We performed a bivariate correlation analysis with the valence/arousal ratings as the first variable, and each of the 13 features as the second variable. We found a significant, strong positive correlation between arousal and five audio features: *pulseclarity* ($r(10) = .79, p < .01$), *zerocross* ($r(10) = .66, p < .05$), *centroid* ($r(10) = .80, p < .01$), *rolloff* ($r(10) = .80, p < .01$), and *brightness* ($r(10) = .73, p < .01$). For valence, apart from a significant, positive correlation with *lowenergy* ($r(10) = .59, p < .05$) and a marginally significant correlation with *mode* ($r(10) = .55, p = .06$), there was no correlation with the remaining audio features.

[1] Ground truth data is available at http://www.ryerson.ca/~nvempala/cmmr2012data.html

3.2 Multiple Regression for Predicting Emotion Ratings

Given that there was some significant correlation between a subset of the 13 audio features and valence/arousal ratings, we performed multiple linear regression to check for a linear relationship between features and ratings. We performed stepwise regression with features as independent variables (probability of F to enter = *.05*) and valence/arousal ratings as dependent variables. The model for arousal, Equation 1, was significant ($F(2,9) = 18.3$, $p < .01$) with *centroid* as the only predictor, accounting for 64.7% of the variance. The model for valence, Equation 2, was significant ($F(2,9) = 5.4$, $p < .05$) with *lowenergy* as the only predictor, accounting for 35.1% of the variance. Here, $y_{Arousal}$ and $y_{Valence}$ are the arousal and valence values on a scale from 1 to 9, respectively. In both equations, the addition of other variables did not lead to an increase in the explained variance. These results clearly suggest that a linear combination of the features does not account well for the valence and arousal ratings of participants. Hence we explored the possibility of predicting valence and arousal ratings through nonlinear combinations of audio features using neural networks.

$$y_{Arousal} = 4.94\, x_{centroid} + 2.14\ . \quad (\mathbf{1})$$

Here, $y_{Arousal}$ is the magnitude of arousal on a scale from 1 to 9.

$$y_{Valence} = 2.12\, x_{lowenergy} + 4.19\ . \quad (\mathbf{2})$$

3.3 Neural Networks for Predicting Emotion Ratings

3.3.1 Static Neural Network

Our first network implementation was a supervised, feedforward network with backpropagation. Our goal was to train the network to predict mean participant valence and arousal values for musical excerpts. We used one set of hidden units for our network. Network architecture consisted of 13 input units, 13 hidden units, and two output units as shown in Figure 1(a). As seen in Table 1, the mean valence/arousal ratings for each of the 12 music excerpts aligned with its expected emotion quadrant. Since valence and arousal ratings were from 1 to 9 and were plotted on the x and y axes respectively, *happy* excerpts needed to have values of $x > 5.0$, $y > 5.0$; *agitated* excerpts needed to have values of $x < 5.0$, $y > 5.0$; *sad* excerpts needed to have values of $x < 5.0$, $y < 5.0$; and *peaceful* excerpts needed to have values of $x > 5.0$, $y < 5.0$. From the 12 music excerpts, we randomly chose two out of three excerpts from each quadrant for our training set, which consisted of M1, M2 (*agitated*), M4, M5 (*happy*), M7, M8 (*peaceful*), and M10, M11 (*sad*). The test set consisted of the remaining four excerpts M3 (*agitated*), M6 (*happy*), M9 (*peaceful*), and M12 (*sad*).

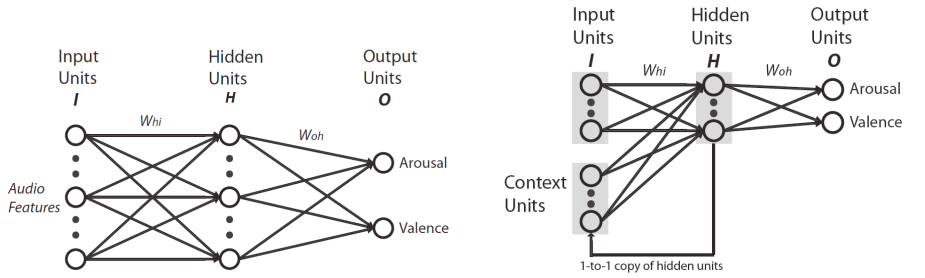


Fig. 1. (a) Static neural network with 13 input units, 13 hidden units, and two output units. W_{hi} indicates connection weights from input units to hidden units. W_{oh} indicates connection weights from hidden units to output units. Only a subset of the 13 input and hidden units are shown. (b) Elman network architecture showing input units, hidden units, output units, and context units. Hidden units from previous processing step are copied into context units for current step.

The network's task was to provide the valence and arousal values based on the 13 audio features. The output values fell within a range of 0 to 1. Since desired outputs were average valence/arousal ratings provided by participants on a scale from 1 to 9, the network outputs were rescaled back. The training set consisted of eight input and output arrays. Each input array had 13 values, one for each audio feature, and its corresponding output array had the two desired arousal and valence values. For example, if the input array being fed into the network was the feature set for excerpt M1 (Bartok), then the input array was [*rms*, *lowenergy*, … , *mode*]= [0.5748, 0.7579, … , 0.0052]^T. Mean participant valence and arousal ratings for M1 were 5.9 and 4.8 respectively, resulting in normalized ratings of 0.61 and 0.48, respectively. Hence the desired output array would be [*arousal*, *valence*]=[0.61, 0.48]^T. To avoid overfitting the network, we kept the number of hidden units equal to the number of input units. The network was built, trained, and tested using the MATLAB programming language. The following procedure was used for training and testing the network:

1. Connection weights W_{hi} (input units to hidden units) and W_{oh} (hidden units to output units) were initialized to random numbers close to zero.
2. Input arrays were fed to the network from the training set in a randomized order. Inputs were passed through a sigmoidal function, multiplied with the connection weights W_{hi}, and summed at each hidden unit. Hidden unit values were obtained by passing the summed value at each hidden unit through a sigmoidal function. These values were multiplied with the connection weights W_{oh}, summed at each output unit, and passed through a sigmoidal function to arrive at the final output value for each output unit. Network outputs were compared to desired outputs and the error was computed. The backpropagation algorithm was applied and changes in connection weights were stored. At the end of the entire epoch, connection weights were updated with the sum of all stored weight changes.
3. The network was trained for approximately 10000 epochs by repeating step 2 to reduce the mean squared error to less than 0.01, and tested. During training, the learning rate parameter was initially set to 0.3 and reduced over time.

We obtained results as shown in Figure 2. The results show the network did a good job of predicting valence/arousal values for M3 (Stravinsky), M9 (Schumann), and M12 (Mozart). Although, predicted values for M6 (Strauss) fell in the expected quadrant (happy), they were not as close to the mean participant ratings. For the purpose of quantifying the network's performance, we computed the Cartesian distance between the mean participant rating and network-predicted value over all four test melodies. The network's performance error was 1.14 on average (on a scale from 1 to 9) or 14.3%, indicating that the network accuracy was 85.7%. These results clearly suggest that a nonlinear relationship exists between music audio features and their associated valence/arousal ratings.

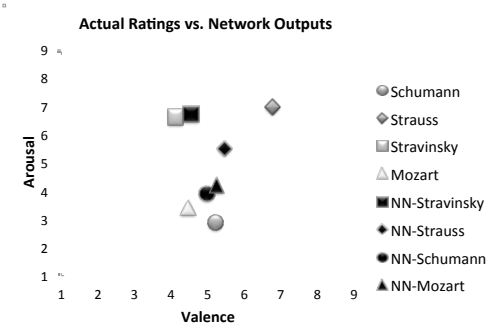


Fig. 2. Mean participant valence/arousal ratings (on a scale of 1 to 9) and corresponding neural network outputs for the four test melody excerpts. *NN* indicates neural network output.

3.3.2 Elman Neural Network for Predicting Emotion Ratings

Although the static network's performance was satisfactory, the network's implementation was based on participant valence/arousal ratings for the entire 120-second duration of each excerpt. It is reasonable to assume that a listener's appraisal of valence and arousal at any point in an excerpt is dynamic and sensitive to the previous few seconds of context. Hence, our next goal was to understand how context might influence a listener's ratings over time. Unlike a typical feedforward network, an Elman network uses context from the previous time-step as additional input for the current time-step. The architecture of the Elman network was almost identical to the static network with 13 input units for features, 13 hidden units, and two outputs units. The network had one additional component, which was a set of 13 context units, as shown in Figure 1(b). Context units were connected to the hidden units similar to input units, and had connection weights associated with them. For each step of input processing in the network, values of hidden units from the previous step were copied to the context units. This is explained below.

Each of the 12 music excerpts was broken into four equal chunks of 30-second duration, and data was collected for the 48 segments from 9 participants, as explained in Section 2. The network was trained on the same 8 music excerpts and tested on the remaining four excerpts, as chosen for the previous network to allow consistent comparison of network performance. The range of input and output values was

identical to the static network. Since the network was being trained on the mean participant data for eight different melodic excerpts, and each excerpt had four 30-second segments, the training set consisted of 32 input and output arrays (four for each excerpt). The four input arrays for each excerpt were sequentially fed into the network. Context units were first initialized to 0. After the input array corresponding to the first segment was processed by the network, values of hidden units were copied into context units for processing the second segment. This process of one-to-one copy from hidden units to context units was continued for the third and fourth segments. This procedure was repeated for all eight excerpts. The network was built, trained, and tested using the MATLAB programming language. The procedure used for training and testing the network was identical to what was used for the static network.

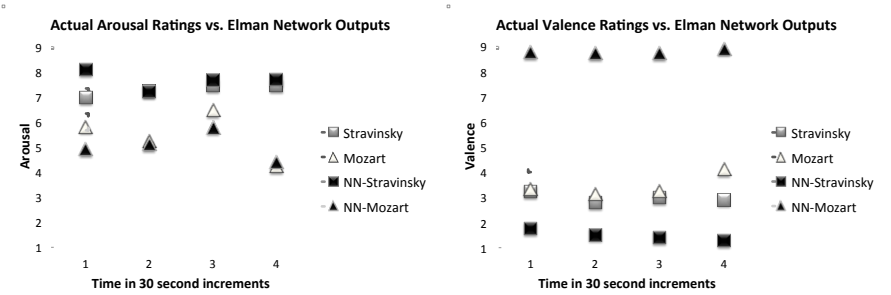


Fig. 4. Mean participant valence (*right*) and arousal (*left*) ratings (on a scale of 1 to 9) and corresponding neural network outputs for the four sequential 30-second segments of two excerpts. *NN* indicates neural network output.

We computed the mean error between participant ratings and network-predicted outputs across all segments of all four test melodies based on Cartesian distance. The network predicted at an average accuracy of 54.3% for all four segments. However, it performed better at predicting valence/arousal values for the final 30-second segment, at an average accuracy of 60%. Figure 4 shows a comparison of mean participant valence/arousal ratings and network values for excerpts M3 and M12. For arousal, the results clearly show that the network was good at predicting how participant ratings were influenced by context from previous segments – i.e., the trend over time was reasonably captured. However, for valence, although the relative changes over time are captured by the network to some extent, the absolute values are poorly predicted.

4 Conclusions and Future Directions

Our aim was to use neural networks to predict valence and arousal ratings of musical excerpts based on audio features within music. Results from the static network indicate that a network can be trained to identify statistical consistencies across audio features abstracted from music and satisfactorily predict valence/arousal values that closely match mean participant ratings. Our second goal was to highlight the role of musical context during listeners' appraisal of emotional content within music, and enable a neural network to utilize previous context during prediction. Results from the

Elman network showed that our network was more successful in capturing the trend of participant appraisals for arousal rather than valence. Three important improvements could be made to our current study involving emotion prediction.

First, having already trained a static network, we would like to identify features that are contributing most towards prediction of valence and arousal. This may be done by removing each feature and testing the network's performance in a step-by-step fashion. Second, a neural network's predictions depend largely on the size and type of training set provided. We intend to train our networks on larger datasets for improved generalizability. We also intend to develop separate static networks that will be trained on different types of musical genres and ratings drawn from different types of music listeners (e.g., trained vs untrained). This would enable us to (a) predict emotion ratings for an excerpt based on its genre and type of listener; and (b) identify salient music audio features for each genre and type of listener. Finally, we would like to improve the performance of our Elman network by training the network on data from a larger set of participants.

Acknowledgments. This research was supported by a Mitacs Elevate postdoctoral fellowship to Naresh N. Vempala co-sponsored by Mitacs and *wave*DNA, Inc. We thank Gillian Sandstrom and Christopher Lachine for assistance with data collection in this study.

References

1. Rainville, P., Bechara, A., Naqvi, N., Damasio, A. R.: Basic Emotions are Associated with Distinct Patterns of Cardiorespiratory Activity. International J. of Psychophysiology. 61, 5--18 (2006)
2. Kim, J., André, E.: Emotion Recognition Based on Physiological Changes in Music Listening. IEEE Transactions on Pattern Analysis and Machine Intelligence. 30, 2067--2083 (2008)
3. Schubert, E.: Modeling Perceived Emotion with Continuous Musical Features. Music Perception. 21, 561--585 (2004)
4. Laurier, C., Lartillot, O., Eerola, T., Toiviainen P.: Exploring Relationships between Audio Features and Emotion in Music. In: 7th Triennial Conference of European Society for the Cognitive Sciences of Music, pp. 260--264. University of Jyväskylä Press, Jyväskylä (2009)
5. Russell, J. A.: A Circumplex Model of Affect. J. of Personality and Social Psychology. 39, 1161--1178 (1980)
6. Laurier, C., Herrera P.: Audio Music Mood Classification using Support Vector Machine. In: Proceedings of ISMIR. Vienna, Austria (2007)
7. Bigand, E., Vieillard, S., Madurell, F., Marozeau, J., Dacquet, A.: Multidimensional Scaling of Emotional Responses to Music: The Effect of Musical Expertise and of the Duration of the Excerpts. Cognition & Emotion. 19, 1113--1139 (2005)
8. Sandstrom, G. M., Russo, F. A.: Music hath charms: The effects of valence and arousal on the regulation of stress. Music and Medicine. 2, 137-143 (2010)
9. Lartillot, O., Toiviainen, P., Eerola, T.: A Matlab Toolbox for Music Information Retrieval. In: Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker, R. (eds.) *Data Analysis, Machine Learning and Applications*, Studies in Classification, Data Analysis, and Knowledge Organization. Springer-Verlag (2008)

Multiple Viewpoint Modeling of North Indian Classical Vocal Compositions

Ajay Srinivasamurthy, Parag Chordia

Georgia Tech Center for Music Technology,
840 McMillan St., Atlanta, USA
{ajays,ppc}@gatech.edu
http://www.gtcmt.gatech.edu

Abstract. Previous research has shown that ensembles of variable length Markov models (VLMMs), known as Multiple Viewpoint Models (MVMs), can be used to predict the continuation of Western tonal melodies, and outperform simpler, fixed-order Markov models. Here we show that this technique can be effectively applied to predicting melodic continuation in North Indian classical music, providing further evidence that MVMs are an effective means for modeling temporal structure in a wide variety of musical systems.

Keywords: Multiple viewpoint modeling, Indian Classical Music, Variable Length Markov Modeling

1 Introduction and Motivation

Melody is an important component in almost all of the world's musical traditions. In North Indian classical music (NICM), it is paramount, and is the basis of a highly sophisticated system of melodic improvisation. Previous work [25] has demonstrated that melodies can be effectively modeled using Multiple Viewpoint Models (MVMs), which are ensembles of variable-length Markov models (VLMMs). Moreover, these models have been shown to accurately reflect listeners expectations about melodic continuation [27]. Chordia [8, 6] generalized this work to tabla sequence prediction; MVMs were shown to be highly effective at modeling the temporal structure of tabla compositions, a percussive tradition based on linear sequences of timbres, suggesting the generality of MVMs for modeling discrete temporal sequences in music. The current work examines whether such models applicable to melodic compositions of NICM. Specifically, we attempt to predict the next note in a symbolically notated melody, given the past context.

2 Background and Related Work

Many aspects of music, such as melodies and chord sequences, can be represented as temporally-ordered sequences of discrete symbols. It is intuitive that how a

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

sequence proceeds will depend most on recent events. For example, melodies are more likely to proceed by small intervals, rather than large jumps [29], which means that the current note constrains the next note. This idea of local dependency can be formalized using Markov models, which are discussed further in Section 5.

A defining characteristic of music is repetition [20]. Subsequences from early in a piece, such as a brief melodic motive, are often repeated later in a piece, sometimes many times. Such repetition also often occurs across pieces; for example, many songs contain common chord sequences. These patterns, which can be short or quite long, present a challenge for a fixed-order Markov models. Although high-order Markov models can be constructed to capture such long-range dependencies, for a length N pattern with K distinct states, there are K^N possible patterns. This means that when learning on a finite set of training sequences, most long patterns will be unseen, leading to extremely sparse transition matrices. A common solution to this problem is to only store sequences that have been seen, replacing a table of counts with a tree structure, called a Prediction Suffix Tree (PST) [28], which is the basis of variable-length Markov models (VLMMs). We describe VLMMs in further detail in Sec. 5. VLMMs form the basis for many music prediction and generation systems [19, 16, 1–3, 22, 18, 17, 15, 24].

In music, there are often multiple ways of representing the musical surface. For example, a melody can be thought of in terms of chromatic pitches, or more abstractly in terms of contour, a sequence of ups and downs. In some cases, patterns may be present when looking at one representation, but not in another. By combining information from multiple viewpoints, it may be possible to capture more of the temporal structure of the sequence. This is the essential idea of MVMs. Each representation, or viewpoint, is modeled using a VLMM and the predictions of each individual model are then combined to compute an overall predictive distribution (Sec. 5). MVMs were introduced by Conklin and Witten [13, 10, 14, 12, 11, 31], and developed by others such as Pearce and Wiggins [26, 25].

3 Indian Classical Music

North Indian Classical Music (NICM) is a centuries-old tradition that is based on melodic and rhythmic improvisation, typically featuring a main melodic instrument or voice and a percussionist. It is organized around *raag*, a melodic abstraction that lies somewhere between a scale, and a fixed melody. A *raag* is most easily explained as a collection of melodic motives, and a technique for developing them. The motives are sequences of notes that are often inflected with various micro-pitch alterations and articulated with an expressive sense of timing. Longer phrases are built by joining these melodic atoms together. Because of this generative process, the musical surface, contains many repeated melodic patterns, making it a natural candidate for modeling with VLMMs.

NICM uses approximately one to two hundred *raags*, of which perhaps fifty are quite common. Although the concept of a note is somewhat different from

that in Western classical, often including subtle pitch motions that are essential rather than ornamental, it is accurate to say that the notes in any given *raag* conform to one of the twelve chromatic pitches of a standard just-intoned scale. It is rare to hear sustained tone that intentionally deviates from one of the twelve chromatic pitches. A given *raag* will use between five and twelve tones.

A typical performance will feature an unmetered elaboration of the *raag* called the *alap* followed by several compositions set to a rhythmic cycle, during which the tabla provides the rhythmic framework. During the rhythmic section there is an alternation between singing the composition and its elaboration, and free improvisations within the context of the *raag*.

Although NICM is largely an oral tradition, in the 20th century there was a push to systematize and notate traditional compositions. The notation that was adopted represented melodies as sequences of discrete notes, having a certain pitch and duration. Additionally certain important ornaments, such as grace notes (*kan swara*) and turns (*khatka*) were indicated as well. Figure 1 gives an example of this notation. Just as with Western music, the notation was not meant to be complete, but to be interpreted within a performance context. In

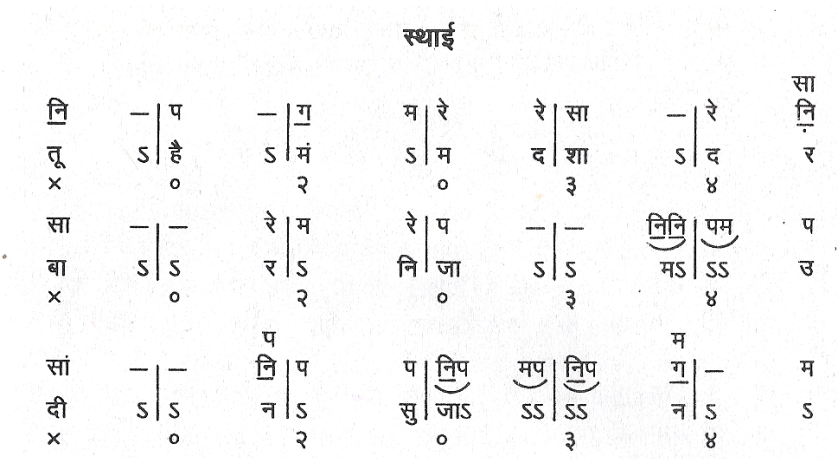


Fig. 1. An example composition of *Raag Suha* in Bhatkhande notation (Vol. II, Page 11 of [21])

this study, we decided to focus on notated compositions because of the difficulty of manual or automatic transcription from audio, which remains for future work.

4 The Indian Classical Music Database

The database used for the study is a part of the NICM symbolic database (bandishDB) being built using *Hindustani Sangeet Paddhati* by Pandit Vishnu

Narayan Bhatkhande [4] and *Abhinav Geetanjai* by Pandit Ramashray Jha [21] which are authoritative works of NICM. The database consists of *bandishes*, vocal compositions with accompanying lyrics. Each *bandish* consists of up to four sections, the *sthayee*, *antara*, *sanchari* and *abhog*. The latter two are relatively rare and are present only in a few compositions. It is worth emphasizing that NICM is largely improvised with the *bandish* providing an initial theme that is heavily elaborated according to the *raag*, within which it is set, and the artist's creativity and virtuosity.

The compositions were first manually encoded into an intermediate text-based representation. Each composition in the symbolic database was then encoded using Humdrum-based syntax called **kern, which was used to encode pitch and duration information. Additionally the following meta-data was stored for each composition: *raag* (melodic framework), *taal* (rhythmic cycle), tempo category (slow, medium, fast). Details of encoding and representation can be found in [5]. An example of the intermediate representation is shown in below.

```
id: jha2011
vol: 2
page: 11
sthayee: Tu hain mammadshah
raag: suha
taal: ektaal
tempo: madhyalaya
//  n  -/  P  -/  g  M/  R  R/  S  -/  R  >Sn,/
//  S  -/  -  R/  M  R/  P  -/  -  nn/  PM  P/
//  S'  -/  -  >Pn/  P  P/  nP  MP/  nP  >Mg/  -  M/
```

Although the artist is free to choose the actual pitch of the tonic in a rendition, all the compositions here are notated with C_4 as the middle tonic. The notes are represented using MIDI note numbers assuming a single key for all the pieces. The notes range from C_3 to B_5 but are folded into one octave from C_4-B_4. The true octave number is stored as an additional parameter. For this study, grace notes and other ornaments such as *meends*(glissandos) were ignored. These ornaments are essential for a complete experience of Indian classical music and the MIDI representation is an approximation, and it does not capture the nuances of singing in its entirety. However, for the present study, a MIDI based representation is sufficient for the analysis of symbolic music scores.

Currently the database consists of 128 compositions in *raags Bageshri, Bihag, Khamaj, Yaman, Yaman Kalyan*, totaling 12,816 notes. When completed, it is expected to be the largest machine readable symbolic NICM database. The data can be freely downloaded at `http://paragchordia.com/data.html`. Table 1 summarizes the dataset used for the experiments in this paper. For this study, compositions from *raags Yaman* and *Yaman Kalyan*, two nearly identical *raags*, are pooled together.

Table 1. Dataset

Raag	***Bandishes***	**Notes**
Yaman	44	4422
Bageshri	36	3720
Khamaj	30	2965
Bihag	18	1709
Total	**128**	**12816**

5 Predictive Modeling

The basic prediction problem can be stated as follows: given a sequence of discretely valued observations, $\{x_1, \ldots, x_{t-1}\}$, compute the next-symbol distribution $P(x_t|x_1, \ldots, x_{t-1})$. In the present case of melody prediction, given the set of symbols (note labels) S, each $x_i \in S$. Given what has occurred so far till time step $t-1$, we wish to predict the next event at time t.

Markov models can be effectively used to model these sequences of symbols, which are often referred to as strings. An n^{th} order Markov model assumes that the next state (associated with a symbol from the alphabet S) depends only on the past n states, i.e. $P(x_t|x_{t-1}, \ldots, x_1) = P(x_t|x_{t-1} \ldots x_{t-n})$. This conditional probability can be calculated by counting how often the symbol x_t follows the context $e_{t-n}^{t-1} = (x_{t-n}, \ldots, x_{t-1})$. Strings of length n are often referred to as n-grams.

Increasing order n, we can model longer strings. However, the number of possible strings ($|S|^n$) increases exponentially. So, even in large databases, most of these n-grams will never be seen, leading to the zero-frequency problem [30, 9]. Variable-length Markov models (VLMMs) address this problem by using an ensemble of fixed-order models, up to order n, to smooth probability estimates. Rather than naively storing counts for all n-grams in a table, to avoid space complexity that increases exponentially with model order, and to make it easy to search for a sequence, n-grams and counts are stored in a partial k-ary tree called a prediction suffix tree (PST) [28]. In the PST, branches represent the succession of certain symbols after others, and a node at a certain level of the PST holds a symbol from the sequence, along with information about the symbol such as the number of times it was seen in the sequence following the symbols above it, and the corresponding probability of occurrence. With this efficient representation of the VLMMs, and with a suitable smoothing method for unseen sequences, we can effectively model melodic sequences.

There are two basic approaches to smoothing: backoff and interpolated. In backoff smoothing, the probability of an unseen sequence is computed by recursively backing off to scaled versions of lower orders. The scale factor applied during each backoff serves as a penalty factor. A backoff smoothing method which adjusts the counts of unseen sequences by adding one, termed as Backoff-A (Method-A in [25]) is explored in this paper. We also explored a simple backoff approach Backoff-B, where there is no penalty for backing off. This allows the

model to back off to lower orders without penalty; the model always chooses a context length for which there is a seen example sequence in the training data.

Interpolation methods compute the probability of a symbol given a context by a weighted interpolation of predicted probabilities at all the orders. A 1/N weighting scheme, as described in [8] is used. For computing probabilities at each order, each node of the PST has some probability mass reserved for unseen sequences, called the escape probability which is added through an extra escape character with a finite escape count. When an unseen sequence is seen in the test sequence at a particular order, the escape probability at that order is returned. We explored interpolated smoothing using an escape count of one at each node (termed as Interp-A) and a very low escape count of 10^{-6} at each node (termed as method Interp-B). Interp-A provides higher escape probabilities while Interp-B method assigns negligible probability mass on unseen sequences. Interp-B is very similar to backing off to lower orders without a penalty, or the Backoff-B method, because of the low escape counts. The smoothing methods and the escape counts define the performance of models, especially at higher orders and with limited training sequence data.

MVMs generalize the idea of combining an ensemble of predictive models. A multiple viewpoints system maintains an ensemble of predictive models based on various viewpoints with varying degrees of specificity. The viewpoints could be basic, derived, or linked viewpoints. Basic viewpoints often refer to the variable being predicted, and are usually observed. Derived viewpoints are obtained from basic viewpoints. Linked viewpoints are cross-type viewpoints which are the Cartesian products of simple and/or derived viewpoints. MVMs finally merge the predictions of these models according to each model's uncertainty at a given time step, using a weighted average as described in [26]. Each viewpoint model is assigned a weight depending on its cross-entropy at each time step. The weight for each model m at time-step t is given by $w_{\mathrm{m}}(t) = H(p_{\max})/H(p_{\mathrm{m}}(t))$, where $H(p_{\mathrm{m}}(t))$ is the entropy of the probability distribution and $H_{\max}(p_{\mathrm{m}})$ is the maximum entropy for a prediction in the distribution. The distributions are then combined by a convex combination, $p(t) = \frac{\sum_m w_m p_m(t)}{\sum_m w_m}$. Higher entropy values result in lower weights. In this way, models that are uncertain (i.e., have higher entropy) make a smaller contribution to the final predictive distribution.

There are two fundamental types of VLMMs which we refer to as long term models (LTMs) and short term models (STMs). LTMs are built on a corpus of songs, while STMs by reading in the symbols, one at a time, from a single composition. The goal of LTMs is to capture patterns that are common across all compositions, while STMs model song-specific patterns. Because songs often contain internal repetition, STMs are often highly predictive. On the other hand, if there is little or no repetition in a song, or the song contains few symbols, LTMs will be more predictive since these have seen much more data. It is also possible to combine the predictions of of the LTM and STM, in a manner analogous to merging viewpoint predictions.

A common domain-independent approach for evaluating the quality of model predictions is cross-entropy [23]. If the true distribution is unknown, the cross en-

tropy of a test sequence of length T can be approximated by $H_c = -\frac{1}{T}\sum_{i=1}^{T}\log_2(p_i)$, which is the mean of the probabilities of true symbols evaluated from predictive distribution at each time step, measured in bits. A closely related concept, often used in natural language modeling is perplexity per symbol [23], defined to be $P = 2^{H_c}$, where H_c is the cross-entropy as described above. Perplexity has a simple interpretation, it is the number of choices that the model is confused between and would be equivalent to the model choosing uniformly between P choices.

6 Experiments

An LTM was built for each *raag*. This was done because the patterns utilized in a bandish are highly dependent on the raag. In future work, it would be straightforward to automatically classify the raag of each bandish, eliminating the need for manually dividing the compositions into *raags* [7].

The viewpoints used in the experiments are listed in Table 2. The viewpoints are obtained from the **kern scores. The note numbers are folded back to lie within the range of a single octave as the absolute notes range over three octaves. The note durations are quantized to the set of Dur={0.125, 0.25, 0.5, 1, 4/3, 2, 8/3, 3, 4, 6, 8}, (where $Dur = 1$ represents the quarter note) in order to limit the total number of duration classes. The melodic interval refers to the interval in semitone between two consecutive notes. This viewpoint, together with the Note viewpoint prevents any loss of information due to the octave folding of notes. The Note⊗Duration viewpoint is a cross-type viewpoint which models the inter-play between pitch and duration.

Table 2. Viewpoints used in the experiments

Viewpoint	Description	Range
Note (N)	The MIDI number of the Note	60, 61,. . ., 71
Contour (C)	A derived viewpoint indicating if the current note is increasing(+1) up, decreasing down(-1) the scale or unchanged(0) from the previous note	-1,0,+1
Interval Change (I)	A derived viewpoint indicating the number of semitones change from the previous note	-11, -10,. . ., 0,. . ., 10, 11
Note⊗Duration(N×D)	A cross-type viewpoint which is 2-tuple of Note and Quantized Duration	$\{(x, y) \mid x \in Note, y \in Dur\}$

For each *Raag*, a leave-one-out cross validation is performed using the compositions from that *Raag*. In each experiment, one composition is chosen as the test composition. The STM is built on the test composition for each viewpoint.

The LTM is trained on the rest of the compositions for each viewpoint. The predictive distribution at each time-step for each viewpoint of LTMs and STMs is computed. The predictions from each viewpoint are merged to obtain combinations such as NI(Note and Interval Change), NC(Note and Contour), NCI(Note, Contour and Interval Change), and NCI+N×D(Note, Contour, Interval Change, and the cross-type N×D). The LTM and STM are also combined into a single predictive distribution as discussed in Sec. 5 and for each case, the perplexity is computed from cross entropy of model predictions. The experiment is also repeated for various different maximum orders of VLMM which correspond to the maximum lengths of symbols modeled at each order.

7 Results and Discussion

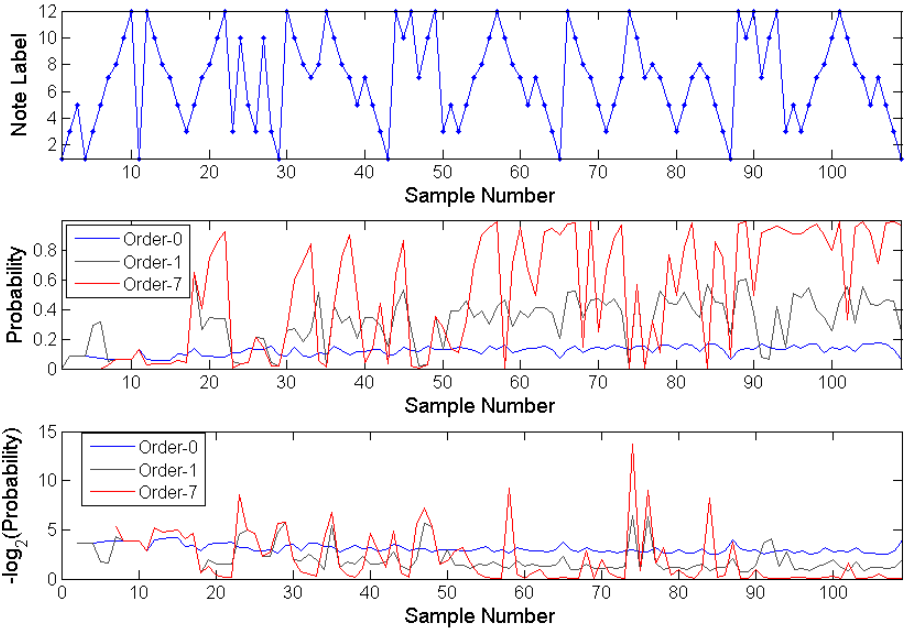


Fig. 2. (a) A single Yaman composition with 109 notes showing distinct repetitive note patterns. (b) The probability p and (c) the log probability, $-\log_2(p)$, of the true symbol as predicted by an STM for different VLMM orders 0 (priors), 1 and 7.

Figure 2 shows a single composition in *Raag Yaman* that contains some repeated motives. We build an STM on the composition and show the probability the model assigned to the true symbol at each timestep for different order STMs. Initially, the probability values are low, but as the composition progresses, the

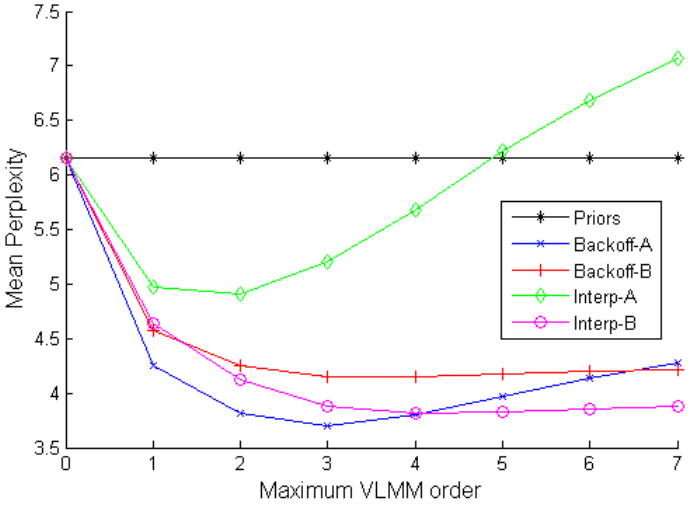


Fig. 3. A comparison of smoothing methods for LTM with the combined viewpoints NCI+N×D

STM is quick to learn the patterns and predicts the true symbol with a high probability. The negative log probability is also shown (in panel (c)), as it is more related to the information in musical events. The peaks in the curve indicate events which are unpredictable. Initially, we see that the negative log probability is higher, and as the piece progresses, the value decreases indicating higher predictability. But there are peaks in the later parts of the piece, which correspond to unexpected changes in note progression. We can also see that there are long term patterns in the piece, which leads to better performance when using higher order models. This demonstrates the effectiveness of STMs in modeling local repetitions. However, these kind of patterns were uncommon in the database – in general, the notated bandishes contained little internal repetition.

In the following discussion, we report results that were found to be statistically significant using a Tukey-Kramer multiple comparison test with confidence bounds at 99%. Figure 3 shows a comparison of smoothing methods for the LTM using the combined viewpoint NCI+N×D. The predictive performance is reported as the mean perplexity of the cross validation experiments, averaged across all the *Raag* models. The priors correspond to probability of true symbols computed through a zeroth order prior distribution of notes through a symbol count. When Interp-A method is used for smoothing, high escape counts lead to flatter predictive distributions with high entropy and hence the perplexity of LTM increases at higher model orders. For reporting further results, we choose the Backoff-A smoothing method which provides a good balance between the probabilities of seen sequences and those of unseen sequences.

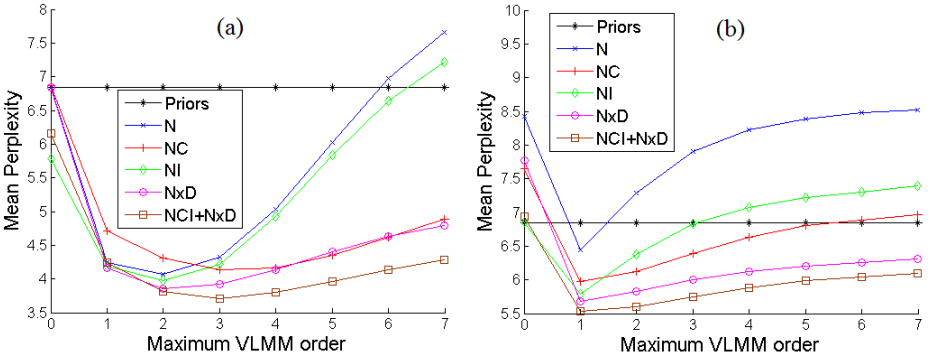


Fig. 4. A comparison of viewpoints with the smoothing method Backoff-A. Panel (a) shows the LTM performance and (b) shows the STM performance

Table 3. The mean perplexity at maximum order of 3 for different *raags* and different viewpoints with LTM and STM. The perplexity of LTM and STM combined using the combination NCI+N×D is also shown. The smoothing method used is Backoff-A.

Raag	LTM						STM					Combined
	Prior	N	NC	NI	N×D	NCI+N×D	N	NC	NI	N×D	NCI+N×D	NCI+N×D
Yaman	6.89	4.58	4.39	4.54	4.14	3.99	8.70	6.80	7.23	6.50	6.05	4.21
Bageshri	6.76	4.15	4.04	4.05	3.72	3.54	7.43	6.11	6.57	5.74	5.54	4.00
Khamaj	7.33	4.48	4.15	4.27	4.17	3.79	7.42	6.16	6.45	5.63	5.55	3.73
Bihag	6.12	3.83	3.67	3.72	3.36	3.18	7.72	6.31	6.98	5.92	5.79	3.47

Figure 4 shows the mean perplexity obtained in the cross validation experiment with different viewpoint combinations, averaged across the *Raags* for both LTM and the STM. Using just the Note viewpoint, the LTM has a lowest perplexity (4.06) when the maximum VLMM order is 2, while we see that combining viewpoints as in NCI+N×D provides the lowest perplexity (3.70) at order 3. Combining viewpoints for both LTM and STM is useful, as the perplexity for the combined decisions are lower than individual viewpoints. This is especially true in LTMs, where the use the multiple viewpoints brings down the perplexity significantly at higher orders. In the case of STMs, the minimum perplexity is consistently achieved at a maximum order of 2. Using the combination of viewpoints NCI+N×D, the perplexity at order-2 drops to 5.54 from the Note viewpoint perplexity of 6.45. The optimal order for LTMs is seen to be 3 while the optimal order for STMs is 1. The low optimal order for STMs show that the training compositions had unpredictable note progressions, even within the same *raag* and hence higher orders are not particularly useful in STMs.

Table 3 consolidates the LTM and STM performance with each *Raag* for different viewpoint combinations. It also tabulates the combined performance of LTM and STM. We see that the model performance is similar across different *Raags*, which indicates that the technique is generalizable to different *Raags*.

The combined performance of LTM and STM is intermediate between LTMs and STMs.

The *bandishes* used in the experiments only provide a basic framework for the actual rendition and lack the repetitions which we normally see in the actual renditions. Further, the *bandishes* are short with about 100 notes per composition. These qualities are reflected in the relatively poor performance of the STMs. Figure 4 shows that the best case perplexity of STM was 1.84 higher than that of the LTM. This is quite different than what was reported in [8], where STMs significantly outperformed LTMs. However the tabla compositions contained, on average 1000 symbols, an order of magnitude more data than the average *bandish*.

The cross entropy of the STM predictions is computed over the entire composition. However, when the STM performance was evaluated only in the second half of the compositions, after the STM has evolved, there was a considerable decrease in perplexity. The smoothing method used contributes to the increasing trend at higher orders. Multiple viewpoints help to reduce the perplexity of melodic prediction and the combination NCI+N×D gives the lowest perplexity.

8 Conclusions and Future Work

MVMs have been shown to effective at predicting melodies in NICM, outperforming fixed and low-order Markov models. This work provides further evidence that MVMs are general tools that can be used to model temporal sequences in a variety of musical genres.

We plan to extend the experiment to include more *raags* once the database is complete. We also plan to extend the work to synthesized audio and develop intermediate-term models based on unsupervised clustering of *bandishes*.

References

1. G. Assayag and S. Dubnov. Universal prediction applied to stylistic music generation. In *Mathematics and Music*, pages 147–160. Springer-Verlag, 2002.
2. G. Assayag and S. Dubnov. Using factor oracles for machine improvisation. *Soft Computing - A Fusion of Foundations, Methodologies and Applications*, 8(9):604–610, 2004.
3. G. Assayag, S. Dubnov, and O. Delerue. Guessing the composer's mind: applying universal prediction to musical style. In *Proceedings of the 1999 International Computer Music Conference*, pages 496–499. San Francisco: ICMA, 1999.
4. V. N. Bhatkhande. *Hindustani Sangeet Paddhati: Kramik Pustak Maalika Vol. I–VI.* Sangeet Karyalaya, 1990.
5. P. Chordia. A system for the analysis and representation of bandishes and gats using humdrum syntax. In *Frontiers of Research in Speech and Music Conference*, 2007.
6. P. Chordia, A. Albin, and A. Sastry. Evaluating multiple viewpoints models of tabla sequences. In *ACM Multimedia Workshop on Music and Machine Learning*, 2010.

7. P. Chordia and A. Rae. Automatic raag classification using pitch–class and pitch–class dyad distributions. In *Proceedings of 8th International Conference on Music Information Retrieval*, Vienna, Austria, 2007.
8. P. Chordia, A. Sastry, and S. Senturk. Predictive tabla modelling using variable length markov and hidden markov models. *Journal of New Music Research*, 40(2):105–118, 2011.
9. J. Cleary and W. Teahan. Experiments on the zero frequency problem. In *Proceedings of Data Compression Conference, DCC '95*, 1995.
10. D. Conklin. Prediction and entropy of music. Master's thesis, University of Calgary (Canada), 1990.
11. D. Conklin. Music generation from statistical models. In *Proceedings of the AISB 2003 Symposium on Artificial Intelligence and Creativity in the Arts and Sciences*, pages 30–35, 2003.
12. D. Conklin and C. Anagnostopoulou. Representation and discovery of multiple viewpoint patterns. In *Proceedings of the 2001 International Computer Music Conference*, pages 479–485. International Computer Music Association, 2001.
13. D. Conklin and J. G. Cleary. Modelling and generating music using multiple viewpoints. In *Proceedings of the First Workshop on AI and Music*, pages 125–137, Menlo Park, CA, 1988. AAAI Press.
14. D. Conklin and I. H. Witten. Multiple viewpoint systems for music prediction. *Journal of New Music Research*, 24:51–73, 1995.
15. A. Cont, S. Dubnov, and G. Assayag. Guidage: A fast audio query guided assemblage. In *Proceedings of International Computer Music Conference (ICMC)*. Copenhagen, September 2007.
16. S. Dubnov. Analysis of musical structure in audio and midi using information rate. In *Proceedings of International Computer Music Conference (ICMC)*, 2006.
17. S. Dubnov, G. Assayag, and A. Cont. Audio oracle: A new algorithm for fast learning of audio structures. In *Proceedings of International Computer Music Conference (ICMC)*, Copenhagen, September 2007.
18. S. Dubnov, G. Assayag, and R. El-Yaniv. Universal classification applied to musical sequences. In *Proceedings of the 1998 International Computer Music Conference*, pages 332–340. San Francisco: ICMA, 1998.
19. S. Dubnov, G. Assayag, O. Lartillot, and G. Bejerano. Using machine-learning methods for musical style modeling. *IEEE Computers*, 36(10):73–80, 2003.
20. D. Huron. *Sweet Anticipation: Music and the Psychology of Expectation.* MIT Press, 2006.
21. R. Jha. *Abhinav Geetanjali Vol. I–V.* Sangeet Sadan Prakashan, 2001.
22. O. Lartillot, S. Dubnov, G. Assayag, and G. Bejerano. Automatic modelling of musical style. In *Proceedings of the 2001 International Computer Music Conference*, pages 447–454. San Francisco: ICMA, 2001.
23. C. Manning and H. Schutze. *Foundations of Statistical Natural Language Processing*, pages 60–78. MIT Press, 2002.
24. F. Pachet. The continuator: Musical interaction with style. In *Proceedings of International Computer Music Conference, Gotheborg (Sweden), ICMA*, pages 211–218, 2002.
25. M. Pearce. *The Construction and Evaluation of Statistical Models of Melodic Structure in Music Perception and Cognition.* PhD thesis, City University, London, 2005.
26. M. Pearce, D. Conklin, and G. Wiggins. *Methods for Combining Statistical Models of Music*, volume 3310, pages 295–312. Springer Berlin, 2005.

27. M. Pearce, M. H. Ruiz, S. Kapasi, G. A. Wiggins, and J. Bhattacharya. Unsupervised statistical learning underpins computational, behavioural and neural manifestations of musical expectation. *NeuroImage*, 50(1):302–313, 2010.
28. J. L. Triviño-Rodriguez and R. Morales-Bueno. Using multiattribute prediction suffix graphs to predict and generate music. *Computer Music Journal*, 25(3):62–79, 2001.
29. P. von Hippel. Redefining pitch proximity: Tessitura and mobility as constraints on melodic intervals. *Music Perception*, 17(3):315–327, 2000.
30. I. H. Witten and T. C. Bell. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory*, 37(4):1085–1094, 1991.
31. I. H. Witten, L. C. Manzara, and D. Conklin. Comparing human and computational models of music prediction. *Computer Music Journal*, 18(1):70–80, 1994.

Comparing Feature-Based Models of Harmony

Martin Rohrmeier[1] and Thore Graepel[2] *

[1] Freie Universität Berlin
[2] Microsoft Research Cambridge
mrohrmeier@cantab.net

Abstract. Predictive processing is a fundamental process in music cognition. While there are a number of predictive models of melodic structure, fewer approaches exist for harmony/chord prediction. This paper compares the predictive performance of n-gram, HMM, autoregressive HMMs as well as feature-based (or multiple-viewpoint) n-gram and Dynamic Bayesian Network Models of harmony, which used a basic set of duration and mode features. The evaluation was performed using a hand-selected corpus of Jazz standards. Multiple-viewpoint n-gram models yield strong results and outperform plain HMM models. However, feature-based DBNs outperform n-gram models and HMMs when incorporating the mode feature, but perform worse when duration is added to the models. Results suggest that the DBNs provide a promising route to modelling tonal harmony.

Keywords: Music; Harmony; Graphical Models; n-gram models; Dynamic Bayesian Networks; Cognitive Modelling; model comparison

1 Introduction

"A mind is fundamentally an anticipator, an expectation-generator." (Dennett, 1996: 57) . Prediction and expectancy formation are fundamental features of our cognitive abilities. The ability to form accurate predictions is important from an evolutionary perspective, ranging from interaction, visual and non-visual perception, synchronisation, complex motor action, or complex communication, be it language or music. Likewise musical expectancy is indispensible for a variety of human musical interactions involving perception, attention and emotion, performance, co-ordination and synchronisation, improvisation, dance. Musical styles ground on established ways to play with patterns of expectancy (such as anticipation, suspense, retardation, revision, garden-path sequences, and deceptive sequences) in order to trigger emotions [20],[11] through the autonomous nervous system based reward and other mechanisms linked with predictive systems [12],[43]. Thus modelling human predictive capacities are fundamental for computational cognitive or interactive models of music.

* We would like to thank Juergen Van Gael, Bob Williamson, Gjergji Kasneci and Marcus Pearce for valuable discussions and ideas.

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

Harmony is one of the core, and at the same time very complex features of Western tonal music. It is a contingent feature of (only) Western music that emerged as an independent structural feature out of modal polyphony around 1650 and constitutes a fundamental feature across nearly all tonal styles, e.g. Classical, Pop, Rock, Latin, or Jazz music. Harmony is governed by the interaction and sequential organisation of chords as discrete musical building blocks that constitute tonality and reflect the formal structure of a piece (such as frequent schemata like the 12-bar blues form or *rhythm changes* in Jazz). In analogy to linguistic syntax, harmony has been argued to exhibit the organisation of a tree structure [15], [14], [45], [42], [5] and was found to be processed in the same (Broca's) area [16]. Since it governs the mid-level organisation of a piece of music, harmony is one of the cornerstones of Western tonal music, notation and its cognition. Hence a rich model of harmony is fundamental for modelling music cognition as well as related music information retrieval tasks such as piece identity/similarity, segmentation, coversong identification, harmonic analysis, searching, indexing, genre classification or generation.

There are plenty of computational models of harmony in the context of computational musical analysis [46],[47] and music information retrieval. This study addresses specifically the problem of harmonic prediction. While much research was done in cognitive studies of melodic and harmonic prediction [48], [43] as well as in modelling melodic prediction (e.g. [30], [29], [26], [27], [13]), comparably little computational work on predictive cognitive modeling of harmony has been done. One successful step to improve predictive n-gram models of melody was the multiple-viewpoint technique [3], [31], which takes different musical features (like duration, onset, scale degree, etc.) into account in order to enhance melodic prediction. This form of feature-based prediction was, however, mostly used for n-gram models and only preliminarily for models of harmony [50]. An approach using similar feature-based HMM or Dynamic Bayesian network models was employed for the problem of automatic chord transcription [19], [28]. DBNs were also successfully employed for note transcription using chords as predictors [33]. Plain n-gram models were used for modelling harmonic structure in a Bach chorale corpus [37], [40], [49], [39], or composer style representation [23]. While different methods for melodic or harmonic n-gram models and representations were compared by [30] and [44], the performance of different types of predictive harmonic models has not been compared. The contribution of this paper is to evaluate plain and feature-based n-gram and graphical models for the prediction of harmony based on a large Jazz corpus.

2 Methods

2.1 Problem setting

Formally, harmonic structure describes a piece of music as a timeseries of discrete, non-overlapping musical building blocks (chords) drawn from a comparably small finite alphabet that feature timing information (metrical structure, onset and duration). The problem of harmony prediction can be expressed as a

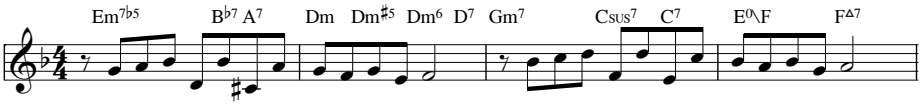
common sequence prediction task for strings of discrete, symbolic events. Given a sequence of events e_i, we let e_a^b denote the subsequence ranging from the indices a to b, employing the notation by [30]. The task is easily specified as modeling the predictive probability distribution $p(e_t|e_1^{t-1})$ of each of the single subsequent events e_t in the chord sequence e_t^T given the sequence of past events e_1^{t-1}. For the current modelling endeavour, we are only focussing on the problem of modelling *which* chord is expected rather than *when* it is expected. As outlined above, the complexity and rich structure of tonal harmony renders it a challenging modeling task which is particularly relevant from a cognitive as well as algorithmic generation perspective.

2.2 Representation

One common problem when modelling harmonic structure is the fact that various forms of harmonic representation are used. In Jazz notation as well as in the formal representation of [9], chords are characterised by root, type, degree attributes and bass note. Such a set of features forms a common denominator shared by different representations (except functional theories, e.g. [36]). According to this representation a chords is represented as specialised sets of pitch classes (roots) with features. A chord symbol such as $E\flat\ aug_G^{7\ 9}$ indicates a (implied fundamental) root, here $E\flat$, a chord type, here aug representing an augmented chord, chord attributes, here 7 9, and a bass note, here G representing the pitch class played in the bass. The richness of this representation and number of chord types is in contrast to frequently employed reduced representations using only 24 major and minor triads (or additional diminished chords), as for instance in the 2008 MIREX task, a fact noted by [19], [44] and others. In the present study a chord was only represented by the Cartesian product of root and type (e.g. $C\sharp m$, $D\flat\ dim$, $G\ aug$) since the other attributes were not sufficiently consistent or correct in the corpus (see below). In addition, we used information about chord duration and mode of the piece.

In total, the data set used here consisted of information drawn from the Cartesian product of chord root, type, duration, and mode (major or minor) of the piece. Due to the considerable computational complexity of the DBN models we refrained from using more additional features in order to motivate this study as a proof-of-concept implementation and a baseline comparison. For the representation of the chord root pitch class, correct enharmonic pitch spelling was used (e.g., $C\sharp \neq D\flat$, $F\flat \neq E\natural$) since this more precise information was available in the corpus and since chord (or pitch) function differs depending on enharmonic pitch spelling and context (for instance, a $E_\flat$ chord in $D\,minor$ may function as tritone substitute of the dominant or Neapolitan chord, if it is a sixth chord, while a $D\sharp$ chord is relatively rare and does not fulfill such functions in $Dminor$). For this purpose we were employing the base-40 representation [10]. The root alphabet consisted of $\{C\flat\flat, C\flat, C, C\sharp, C\sharp\sharp, D\flat\flat, ...\}$. The chord type alphabet occuring in the corpus was $\{maj, min, dim, hdim, aug, alt, sus\}$. Two additional padding symbols in the alphabet denoted the beginning and end of a piece. Further, chord duration (dur) was represented in beats, and $mode$ (of the

Fig. 1. Representation of harmony exemplified by the standard "You must believe in Spring" (above in Real Book style). The top part of the table represents the (relevant) harmony information coded in the corpus. Features used for the model comparison are marked by an asterisk.



Chord	Em7b5	Bb7	A7	Dm	*Dm*♯5	Dm6	D7	Gm7	C7sus	C7	Edim/F	FMaj7
bar	1	1	1	2	2	2	2	3	3	3	4	4
beat	0	2	3	0	1	2	3	0	2	3	0	2
root*	E	Bb	A	D	D	D	D	G	C	C	e	F
type*	hdim	maj	maj	min	min	min	maj	min	sus	maj	dim	maj
bass	E	Bb	A	D	D	D	D	G	C	C	F	F
att	(7b5)	7	7	-	♯5	6	7	7	7	7	-	M7
chord*	E hdim	Bb maj	A maj	D min	*Dmin*	D min	D maj	G min	C sus	C maj	E hdim	F maj
dur*	2	1	1	1	1	1	1	2	1	1	2	2

piece) as binary variable (*major / minor*). Altogether an alphabet of 135 chord symbols occurs in the corpus.

2.3 Models

According to the considerations outlined above, we compared three types of models, (i) multiple viewpoint n-gram models that have been very successfully used (mostly for modelling melody), (ii) HMMs which were also commonly used for musical applications. Further we propose (iii) a novel type of graphical model extending the HMMs by a feature-based approach in analogy to the n-gram methodology.

Multiple Viewpoint n-Gram Models Multiple viewpoint n-gram models (feature-based models) were first suggested for the application to music, and in particular to melodic structure, by [3]. The methodology was extended and extensively evaluated by [29]. It constitutes the heart of the information dynamics of music model (IDyOM, www.idyom.org). The idea behind the multiple-viewpoint technique (MVP) is to combine n-gram models for different structural features, in this case features such as duration, metrical structure or mode (of the piece). Combined viewpoint models (such as $chord \otimes mode$) project the prediction space down to the viewpoint to be predicted by means of marginalising over the other viewpoints (see [29],[3] for details). Such n-gram models have to avoid zero counts and to gain confident predictions between the extremes of overfitting (based on too large contexts) and using overly unspecific information (from too short contexts). A large-scale comparison of different smoothing and interpolation techniques [30] found that Witten-Bell smoothing [51], [17] performed best for the case of melodic prediction. Our implementation of harmonic feature-based n-gram models employed Witten-Bell smoothing and mode, duration, chord, root and type viewpoints.

We use $\kappa_{i,n}$ as a shorthand for the predictive context of the n-gram model, i.e. subsequence $e_{(i-n)+1}^{i-1}$. The probability distribution $\hat{p}(e_i|\kappa_{i,n})$ of the predictive event is modelled by the weighted sum of predictions of all available context-lengths (2). The probability $\alpha(\kappa_{i,n})$ is approximated by the count $c(\kappa_{i,n})$ of n-grams of the given context $\kappa_{i,n}$ and adding a zero-escape approximation to the denominator, which amounts to the number of the encountered symbol types $t(\kappa_{i,n})$. The number of encountered symbol types $t(\kappa_{i,n})$ is further used to adjust the weight of the escape count $\gamma(e_i^j)$ to be approximately proportional to the number of symbol types. ζ denotes the alphabet of surface symbols.

$$\kappa_{i,n} := e_{(i-n)+1}^{i-1} \tag{1}$$

$$\hat{p}(e_i|\kappa_{i,n}) = \alpha(e_i|\kappa_{i,n}) + \gamma(\kappa_{i,n})\ \hat{p}(e_i|\kappa_{i,n-1}) \tag{2}$$

$$\gamma(\kappa_{i,n}) = \frac{t(\kappa_{i,n})}{\sum_{c\in\zeta} c(\kappa_{i,n}) + t(\kappa_{i,n})} \tag{3}$$

$$\alpha(\kappa_{i,n}) = \frac{c(\kappa_{i,n})}{\sum_{e\in\zeta} c(e|\kappa_{i,n}) + t(\kappa_{i,n})} \tag{4}$$

Hidden Markov Models Hidden Markov Models (HMM, [32]) are well-known and do not require a detailed introduction. They are successfully applied across domains, including melodic models, harmonisation, harmonic labelling, transcription or audio alignment problems [35], [34], [1], yet not extensively in contexts that involve harmonic prediction. An HMM models a discrete symbol sequence as a series of symbol emissions the distributions of which are controlled by an underlying Markov process representing a number of hidden states by a single discrete random variable. Inference is performed based on maximum likelihood estimate using the Baum-Welch algorithm. The likelihood of a sequence and prediction is computed based on the forward algorithm [32], [22]. We used the implementation provided by Kevin Murphy's Bayes Net Toolbox (BNT,[21]).

Feature-Based Dynamic Bayesian Networks While HMMs and related graphical models have been employed for the case of music and modelling the relationship of notes and chords [24], [25], [27], [28], we employ a graphical generalisation of the multiple-viewpoint idea that combines and to generalise the idea of viewpoints/ feature-based prediction as applied successfully in n-gram models with the flexibility of greater sequential contexts as available to HMM models. This way the model makes use of the hidden state space as well as principled inference over features compared to the heuristic blending used in n-gram models. We build on *state-space models* and Murphy's research on *Dynamic Bayesian Networks* [22]. For applications of chord transcription, [19] used a harmonic DBN conditioning on key and metrical structure as higher-order model in his transcription system. [28] integrated a representation of metre into the transition matrix of an extended HMM model. In our suggested architecture we make use of *mode* and *duration* features such that the current hidden state depends not

only on the previous state but also on mode and/or previous duration (see Figure 2). We further implemented auto-regressive versions of these models which combine the feature-based prediction with conditioning on the previous chord, and hence incorporate the predictive power of bigram models (cf., [22]). The models were implemented using the unrolled *junction tree* inference algorithm within the BNT framework.

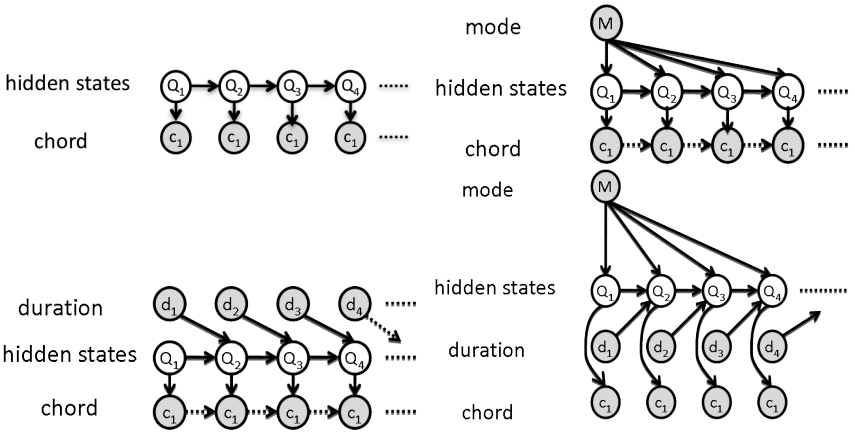


Fig. 2. Architecture of the four types of Dynamic Bayesian Networks, unrolled for 4 time steps. The figure displays the plain Hidden Markov Model (top left), and the structure of its feature-based DBN generalisations using either *mode* (top right), *duration* (bottom left) or both (bottom right). The dotted arrows represent additional auto-regressive versions of these models.

2.4 Evaluation

The corpus consisted of 1,631 hand-selected Jazz pieces from the Band-in-a-Box (BiaB) corpus [6]. Since the original BiaB corpus contained numerous community-entered pieces which in part contain hundreds of orthographic and syntactic mistakes, the selection was made by a human Jazz expert. The prepared corpus was divided into a *training set* of 1,471 pieces, which featured 107,505 chords, and a *testing set* of 160 pieces. The database contained the BiaB metainformation tags, full chord information, chord onset, chord duration as well as song structure. The dataset was preprared from the original BiaB data format based on Mauch's MGU-format reader [18]. The dataset is available online at `www.mus.cam.ac.uk/CMS/people/mr397/`.

For the evaluation, models were trained on the training set and subsequently evaluated on the testing set by predicting each chord of each individual piece. Cross-entropy H_m and perplexity PP_m measures (5, 6) were averaged across the corpus.

$$H_m(p_m, e_1^T) = -\frac{1}{T}\sum_{t=1}^{T} \log_2 p_m(e_t|e_1^{t-1}) \tag{5}$$

$$PP_m(p_m, e_1^T) = 2^{H_m(p_m, e_1^T)} \tag{6}$$

3 Results and Discussion

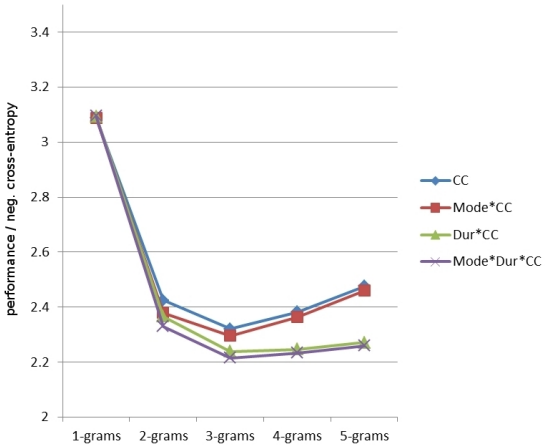
3.1 Feature-Based n-Gram Models

Figure 3 displays the results. The performance of the simple feature-less n-gram models (operating only on the core chord) illustrates that trigram models perform best, while higher-order models tend to overfit the data. This confirms the findings by [30] for melody in the domain of harmony. In order to have a base-line estimate for the model performance under the impact of knowledge of particular, idiosyncratic sequences (modelling "veridic" tonal knowledge or the partial impact of multiple listening [2], [8]) performance was also computed having the training set in the testing set (also done by [44]). The increase in performance for the veridic evaluations shows that the impact of idiosyncratic sequences is high in 4-grams while 5-gram models are close to be optimal. Perplexity results are similar to/augment results based on cross-entropy. Under all conditions, 3-gram models perform best, while higher-order models tend to overfit. The inclusion of features/viewpoints yields significant improvements over blank n-gram models.[3]

The improvement when adding *mode* is relatively small. However, the duration viewpoint (adding *dur* and $mode \otimes dur$) improves performance remarkably. The mode feature improves performance only slightly. The results thus indicate that n-grams contain mode information to a large extent. This is plausible given that single chords or chord progressions distinguish both modes (e.g. $Dm^{7\flat5}$ is unambiguous $\hat{2}$ of *C minor* while *Dm* identifies $\hat{2}$ in *C major*, *Fm G* is an unambiguous minor progression). This is consistent with the distinct differences between the top major or minor harmonic n-gram frequencies identified in Bach's chorales [37]. Thus the additional feature does not yield much further enhancement (this explanation is further underpinned by the fact that the improvement of including mode is larger for bigram than for trigram oder higher-order models). In analogy, the combined feature $mode \otimes duration$ yields best performance and improves the performance for the duration feature only slightly. The reason why duration improved chord prediction significantly may be that duration is an indicator of chord stability: for instance, shorter chords may occur ornamentally with respect to other chords [42] and, in consequence, behave differently in context than more stable, longer chords.

[3] Preliminary tests found (unsurprisingly) that the prediction accuracy is considerably better for the Cartesian product representation of core chords than the viewpoint-based combination of its components (e.g. $root \otimes type$). This confirms the music theoretical notion that the core components root and type are strongly dependent, an effect that cannot be captured by treating them as independent viewpoints.

Fig. 3. N-gram performance results for different plain and feature-based models.



	chord	veridic *chord*	*mode* ⊗ *chord*	*dur* ⊗ *chord*	*mode* ⊗ *dur* ⊗ *chord*
1-grams	3.09	3.08	3.09	3.09	3.10
2-grams	2.42	2.36	2.38	2.37	2.33
3-grams	**2.32**	1.97	**2.30**	**2.24**	**2.22**
4-grams	2.38	1.54	2.36	2.25	2.23
5-grams	2.48	1.11	2.46	2.27	2.26

3.2 Hidden Markov Models

Figure 4 displays the performance of the HMM model dependent on the number of hidden states. The HMM performance reaches its best performance at around 2.47 (negative cross-entropy) using 65 hidden states. The mean performance reaching from 25 to 150 hidden states is, however, 2.49. This indicates that a large range of models performes at a similar level and that it infers a comparable amount of information about the chord sequences, independently of whether the inferred information represented in the prior, hidden states, and emission vectors reflects music theoretically established distinctions or not. A large number of hidden states does not improve the performance of the model and extract further harmonic information. With respect to the overfitting problem, HMMs of increasing complexity do not exhibit a comparably strong effect of overfitting as found with n-gram models of increasing context-length: as figure 4 illustrates, the performance decreases comparably slowly for larger numbers of hidden states. Hence HMMs are to some extent more robust with respect to the problem of overfitting. The extent to which the model structure reflects human theoretical knowledge remains to be addressed in further research.

The overall predictive power of HMM models is worse compared with n-gram models. Even the best performing HMMs rank lower than 2-bigram models.

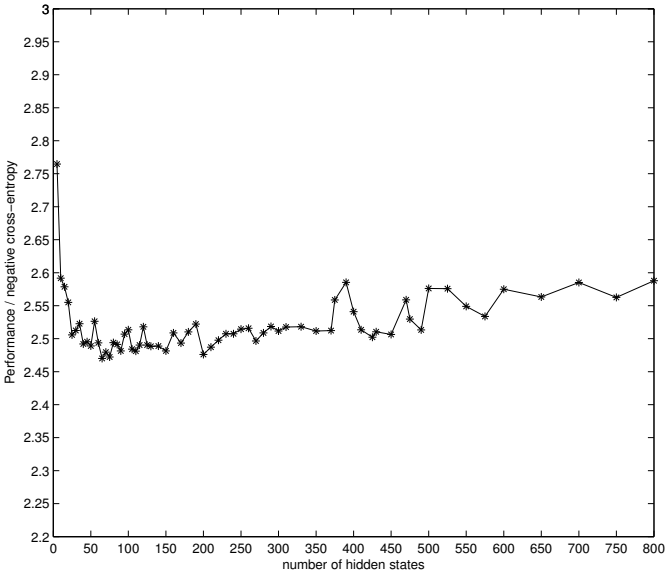


Fig. 4. Performance of plain HMMs.

This finding matches with common results that n-gram models generally tend to outperform other types of models when it comes to prediction accuracy (cf. [17]). The poor performance of HMMs motivates the exploration of enhanced feature-based graphical models in order to improve the predictive power.

3.3 Feature-Based Dynamic Bayesian Networks

The intention behind the feature-based DBNs was to combine the strength of HMM-based sequence modeling with incorporating additional feature information. However, such extensions increase the model complexity considerably and hence, only a small fraction of the design space of these types of models could be explored within reasonable time constraints. Since such graphical viewpoint models have not been evaluated for harmony prediction, the current results constitute a proof of concept and provide estimates of their performance. Table 1 displays the results for the different types of candidate architectures. Every time chord symbols were encoded as Cartesian products of their components because preliminary results showed that models in which these features were separated performed worse.

As expected, auto-regressive HMMs (without viewpoints) perform better than bigram models even with small numbers of states. This confirms that the autoregressive conditioning on the previous chord symbol is (at least) equivalent

Table 1. Results for Hidden Markov Models and Dynamic Bayesian Networks incorporating different feature combinations, numbers of hidden states (hid) and optimal HMM performance for reference. Performance is represented in terms of negative cross-entropy and perplexity.

HMM			**DBNs**								
chord			*mode* ⊗ *chord*			*dur* ⊗ *chord*			*dur* ⊗ *mode* ⊗ *chord*		
hid	-CE	PP	hid	-CE	PP	hid	-CE	PP	hid	-CE	PP
30	2.51	6.12	30	2.42	5.79	30	2.36	5.57	30	**2.35**	5.65
90	2.48	5.90	90	2.28	5.34	90	**2.29**	5.55	130	2.41	7.48
130	2.48	5.96	130	**2.26**	5.45						
best											
65	**2.47**	5.91									

Auto-regressive models								
chord			*mode* ⊗ *chord*			*dur* ⊗ *chord*		
hid	-CE	PP	hid	-CE	PP	hid	-CE	PP
10	**2.36**	5.53	10	2.28	5.51	15	2.26	5.50
15	2.36	5.53						
25	2.36	5.53						
50	2.36	5.53						

to a bigram model. The better performance indicates that they encode relevant information in the hidden states, yet to a limited extent: the performance does not increase as the number of hidden states increases and they do not reach the performance of 3-gram models. Accordingly, this extension is found not to be advantageous for the HMM.

In contrast, when mode is added as a feature, the DBN outperforms both the mode n-gram models as well as the plain HMM and improves performance with increasing number of states. This indicates that the adding of the mode feature strongly increases the performance and that the Dynamic Bayesian Network extracts further distinctive features of major and minor modes than both other model types. Again this may imply that the model draws information from the longer available context. In contrast to the plain models, the auto-regressive extension of this DBN in turn does not yield additional improvement. Yet conditioning on mode renders the feature-based DBN the best performing model for this feature.

For the present sample computations, the additional duration feature conditioning on the hidden states raises the model complexity drastically but does not yield an improvement of performance compared with the mode feature or the plain HMM (the improvement achieved by auto-regressive conditioning, however, provides a hint that this complex type of model performs as good as the (best) yet simpler mode-featured DBN). Nonetheless none of the DBN models can make comparably efficient use of the additional information embodied in chord durations as multiple-viewpoint n-gram models (reaching a performance of 2.24 and 2.22, see Figure 3). We assume that the comparably weak performance of the

model was due to the fact, that a proportion of the duration values was sparse and that the implementation based on the BNT toolbox lacked specific methods to deal with this problem (to be addressed in future model development). Not surprisingly, the DBN model which combined both mode and duration information performed at the level of the respective multiple-viewpoint bigram models, its performance was, however, much lower than the trigram model performance. This is likely to be due to the same sparsity problem of the distribution of duration values.

Altogether the results show that the inclusion of additional feature information enhances the prediction (and modelling) of harmony. When only chord information was used, n-gram models strongly outperformed HMM models. Autoregressive HMMs, however, achieved a performance improvement over bigram models. When using all available information (chord, mode and duration), n-gram models performed best, which may be explained by the application of enhanced smoothing methods dealing with data sparsity in duration values. Using only mode information, DBN and autoregressive DBN models performed best. This suggests that the combination of combining feature information (without sparsity) and the longer available context yields predictive performance better than n-gram models. This finding suggests that the further development of more advanced feature-based DBN models may provide a promising flexible type of a graphical predictive, cognitive model.

4 Conclusion

The paper presented a comparison of a set of feature-based n-gram, HMM and graphical feature-based Dynamic Bayesian Network models with repsect to predictive modelling of complex harmonic structure in music. These models are important for computational, cognitive and descriptive approaches to music as well as practical applications in terms of generation or real-time interaction. A model comparison showed harmonic prediction improved taking mode or duration information into account. The improvement using mode, however, was comparably small reflecting that differences between both modes are to an extent already reflected in short chord fragments [37]; thus incorporating a dynamic mode feature may be expected to yield only small further improvment. This underpins that harmonic structure is governed not only on mere chord information but also temporal, key or potentially other features. Multiple-viewpoint n-gram models [3],[29] produced best results when duration information is utilised. When only using mode information, however, they are outperformed by feature-based DBNs. These proof-of-concept evaluations illustrate that feature-based DBNs combining the feature approach with the HMM architecture constitutes a promising avenue for further predictive and cognitive modelling of harmony. However, in order to arrive at rich cognitive models further refinements of feature-based model types are required in order to incorporate inference and prediction of metrical structure, dynamic features (like key or mode changes) as well as higher-order features like scale-degree or tonal function (similarly to higher-order melodic viewpoints,

[3], [29]). Moreover, from a theoretical perspective the organisation of harmonic sequences was argued to be hierarchical and exceed simple local Markovian dependencies [45], [38], [40], [42], which in turn may suggest that computational models of similar complexities would yield further improvement.

Ultimately the cognitive aspect of the prediction task requires human baseline measures since the criterion of optimality for the computational models is not most accurate prediction but similar behaviour as human minds – otherwise, intentional musical effects of unpredictable structures like harmonic garden-path or surprise sequences could not be modeled. Future comparison with human experimental results from priming or event-related potential studies may yield further insights into to cognitive adequacy of such probabilistic models and human predictive information processing [43].

References

1. Allan, M., Williams, C. K. I.: Harmonising chorales by probabilistic inference. Advances in Neural Information Processing Systems 17 (2005)
2. Bharucha, J.J., Krumhansl, C.L.: The representation of harmonic structure in music: Hierarchies of stability as a function of context. Cognition 13, 63–102 (1983)
3. Conklin, D., Witten, I.: Multiple viewpoint systems for music prediction. Journal of New Music Research 24(1), 51–73 (1995)
4. Dahlhaus, C.: Untersuchungen über die Entstehung der harmonischen Tonalität. Bärenreiter, Kassel (1967)
5. De Haas, W.B., Rohrmeier, M., Veltkamp, R., Wiering, F.: Modeling harmonic similarity using a generative grammar of tonal harmony. In: 10th International Society for Music Information Retrieval Conference (ISMIR 2009), pp. 549–554 (2009)
6. DeHaas, W.B., Veltkamp, R., Wiering, F.: Tonal pitch step distance: a similarity measure for chord progressions. In: 8th International Conference on Music Information Retrieval (ISMIR) (2008)
7. Dennett, D.: Kinds of Minds. Basic Books, New York (1996)
8. Eerola, T.: The Dynamics of Musical Expectancy. Cross-Cultural and Statistical Approaches to Melodic Expectations. PhD thesis, University of Jyväskylä (2003)
9. Harte, C., Sandler, M., Abdallah, S.A., Gmez, E.: Symbolic representation of musical chords: A proposed syntax for text annotations. In: Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR), pp. 66–71 (2005)
10. Hewlett, W.B.: A base-40 number-line representation of musical pitch notation. Musikometrika 4,1–14 (1992)
11. Huron, D.: Sweet Anticipation: Music and the Psychology of Expectation. MIT Press, Cambridge, Massachusetts (2006)
12. Koelsch, S.: Towards a neural basis of music-evoked emotions. Trends in Cognitive Sciences 14(3), 131–137 (2010)
13. Lavrenko, V., Pickens, J.: Polyphonic music modeling with random fields. In: Proceedings of ACM Multimedia, Berkeley, CA (2003)
14. Lerdahl, F.: Tonal Pitch Space, Oxford University Press, New York (2001)
15. Lerdahl, F. and Jackendoff, R.: A Generative Theory of Tonal Music. MIT Press, Cambridge, MA (1983)
16. Maess, B., Koelsch, S., Gunter, T.C., Friederici, A.D.: Musical syntax is processed in broca's area: An meg study. Nature Neuroscience 4(5):540–545 (2001)

17. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA:, 1999.
18. Mauch, M., Dixon, S., Harte, S., Casey, M., Fields, B.: Discovering chord idioms through Beatles and Real Book songs. In: 8th International Conference on Music Information Retrieval (ISMIR) (2007)
19. Mauch, M., Dixon, S.: Simultaneous estimation of chords and musical context from audio. IEEE Transactions on Audio, Speech, and Language Processing 18(6),1280–1289 (2010)
20. Meyer, L.B.: Emotion and Meaning in Music. University of Chicago Press, London (1956)
21. Murphy, K.: The bayes net toolbox for matlab. Computing Science and Statistics 33 (2001).
22. Murphy, K.: Dynamic Bayesian Networks: Representation, Inference and Learning. PhD thesis, UC Berkeley (2002)
23. Ogihara, M., Li, T.: N-Gram chord profiles for composer style identification. 9th International Conference on Music Information Retrieval (ISMIR), pp. 671–676 (2008)
24. Paiement, J.F., Eck, D., Bengio, S.: A probabilistic model for chord progressions. In 6th International Conference on Music Information Retrieval (ISMIR) (2005)
25. Paiement, J.F.: Probabilistic Models for Music. PhD thesis, Ecole Polytechnique Fdrale de Lausanne (2008).
26. Paiement, J.F., Bengio, S., Eck, D.: Probabilistic models for melodic prediction. Artificial Intelligence 173(14), 1266–1274 (2009)
27. Paiement, J.F., Grandvalet, Y., Bengio, S.: Predictive models for music. Connection Science 21(2-3), 253–272 (2009)
28. Papadopoulos, H., Peeters, G.: Joint estimation of chords and downbeats from an audio signal. IEEE Transactions on Audio, Speech, and Language Processing 19(1),138–152 (2011)
29. Pearce, M.T.: The construction and evaluation of statistical models of melodic structure in music perception and composition. PhD thesis, City University, London(2005)
30. Pearce, M.T., Wiggins, G.A.: Improved methods for statistical modelling of monophonic music. Journal of New Music Research 33(4), 367–385 (2004)
31. Pearce, M.T., Wiggins, G.A.: Expectation in melody: The influence of context and learning. Music Perception 23(5), 377–405 (2006)
32. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE 77:257–286 (1989)
33. Racyński, S., Vincent, E., Bimbot, F., Sagayama, S.: Multiple pitch transcription using DBN-based musicological models. In: 10th International Conference on Music Information Retrieval (ISMIR) (2009)
34. Raphael, C.: Music plus one and machine learning. In: Machine Learning, 27th International Conference (ICML), (2010)
35. Raphael, C., Stoddard, J.: Functional analysis using probabilistic models. Computer Music Journal 28(3), 45–52 (2004)
36. Riemann, H.: Vereinfachte Harmonielehre; oder, die Lehre von den tonalen Funktionen der Akkorde. Augener (1893)
37. Rohrmeier, M.: Towards modelling movement in music: Analysing properties and dynamic aspects of pc set sequences in Bach's chorales. Darwin College Research Reports 04, University of Cambridge, `www.darwin.cam.ac.uk/dcrr/` (2006)
38. Rohrmeier, M.: A generative grammar approach to diatonic harmonic structure. In: Spyridis et al. (eds.) 4th Sound and Music Computing Conference, pp. 97–100 (2007)

39. Rohrmeier, M.: Modelling dynamics of key induction in harmony progressions. In: Spyridis et al. (eds.) 4th Sound and Music Computing Conference, pp. 82–89 (2007)
40. Rohrmeier, M., Cross, I.: Statistical Properties of Harmony in Bach's Chorales. 10th International Conference on Music Perception and Cognition (ICMPC), pp. 619–627 (2008)
41. Rohrmeier, M.: Implicit learning of musical structure: Experimental and compuational modelling approaches. PhD thesis, University of Cambridge (2010)
42. Rohrmeier, M.: Towards a generative syntax of tonal harmony. Journal of Mathematics and Music 5(1), 35–53 (2011)
43. Rohrmeier, M., Koelsch, S.: Predictive information processing in music cognition. a critical review. International Journal of Psychophysiology 38(2), 164–175 (2012)
44. Scholz, R., Vincent, E., Bimbot, F.: Robust modelling of musical chord sequences using probabilistic n-grams. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 53–56 (2009)
45. Steedman, M.J.: The blues and the abstract truth: Music and mental models. In: Garnham, A., Oakhill, J. (eds.) Mental models in cognitive science, pp. 305–318. Erlbaum, Mahwah, NJ (1996)
46. Temperley, D.: The Cognition of Basic Musical Structures, MIT Press, Cambridge, MA (2001)
47. Temperley, D.: Music and probability, MIT Press, Cambridge, MA (2007)
48. Tillmann, B.: Implicit investigations of tonal knowledge in nonmusician listeners. Annals of the New York Academy of Science 1060, 100–110 (2005)
49. Tymoczko, D.: Function Theories: A Statistical Approach. Musurgia 10 (3-4), 35–64 (2003)
50. Whorley, R., Wiggins, G.A., Rhodes, C.S., Pearce, M.T.: Development of techniques for the computational modelling of harmony. In: Ventura et al. (eds.) International Conference on Computational Creativity, Lisbon (2010)
51. Witten, I.H., Bell, T.C.: The zero-frequency problem: estimating the probabilities of novel events in adaptive text compression. IEEE Transactions on Information Theory 37(4), 1085–1094 (1991)

Music Listening as Information Processing

Eliot Handelman[1] and Andie Sigler[1,2]

[1] Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT)
[2] School of Computer Science, McGill University
andrea.sigler@mail.mcgill.ca

Abstract. Computational reasoning about musical structure (in particular, pattern, shape, and motion), with a perceptual and mathematical basis in simplicity.

Keywords: Musical structure, artificial intelligence, computer models, simplicity

1 An Information Processing Task

Paralleling a theory of vision (as stated e.g. by Marr [1]), it is possible to treat musical listening as an information processing task. In computational vision research, input is known as "I" for image, and the goal is to find functions that elucidate various properties of this image. The dual to this can be found in the (drawing-automaton) work of Harold Cohen [2], who asks "what is the minimum condition under which a set of marks functions as an image?" Both questions are also fundamental questions for music research, with the "images" in question being musical instead of visual.

Marr's three levels of description for an information processing task are the computational, the algorithmic, and the hardware. At the computational level, we can ask what is the *function* (or program) to be computed; at the algorithmic level we ask what are the *types* or the *language* in which such a function may be expressed; at the hardware level, we ask how functions in the language can be run on a physical computer (or brain).

This paper is focused at the algorithmic level, on the development of a few *types* for reasoning about musical structure.[3] The goal is to facilitate a higher level of structural description, so that further studies at the computational level (these may be statistical, musicological, generative, etc.) may be carried out on multi-dimensional complexes of pattern and shape, rather than on sequences of notes.

It is useful for a system of types to be *composable*. A composable system is one in which higher-level types can always be built from lower-level types

[3] These types have been implemented as part of a large-scale musical AI project, and can be explored through an interactive visualization system. Promising experiments have been made towards using types to direct "orchestrations" in order to sonify aspects of an analysis.

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

without having to define new types or composition functions at each new level. For example, given a type *list* and a function to put (any) things together into a list, we automatically get *lists of lists of lists of lists* and so on.

The types proposed for describing musical structure fall into the three categories of *pattern*, *shape*, and *motion*.[4] These types are composable, allowing us to speak of a pattern of patterns, a shaped and patterned motion, a pattern that has a shape, and so on.[5] With just a few definitions, therefore, it is possible to build from local structure up to the large scale.

In order for types to be composable, it is necessary that they be *rational* (i.e. logical or algebraic). Therefore, commonplace music-theoretic notions such as the labeling of harmonies, keys, or typical forms are avoided. Any system of labeling that is non-exhaustive is an *empirical* system, based on knowledge (and theorization) of existing music. The rational methodology we propose is exhaustive in the sense that we can't "look for" ABA patterns without also seeing AAA, ABB, ABC, $ABCD$, and so on. In the computational study of music, it makes sense for a rational level to come first. Later studies may impose or deduce an empirical level on top of the basic rational level, perhaps by constraining general types to fit a particular theory or answer a particular question, or by using statistical techniques on musical corpora.

Since questions of style and musical "language" are empirical, the utility of a system of rational types should not depend on the kind of music to which they're applied. At a rational level of structural analysis, we *do not need to know* the context or genre of the music in advance of the analysis.

What is required is a theorization of basic musical material and its possibilities. For example, we assume nothing of a melody but that it's a succession of pitches, each with a duration, and that pitch may be represented as a total ordering (i.e. for two pitches X and Y, there are exactly three options: X is higher than Y, Y is higher than X, or they are equal). Given an input "image" of this sort, what kinds of structures are inevitably found within?

1.1 Simplicity

On a strictly mathematical basis, there are many possible answers to the foregoing question. To narrow the field, we can ask "what kinds of structures might be interesting to explore?"

Since music is made by and for the human mind, an interesting analysis might recognize the kinds of structures that would be most *obvious* to a human: *simple* structures. Simplicity is interesting because in music, as elsewhere, the simplest things are the most salient (obvious), and certainly all music distinguishes within its discourses differing degrees of salience. Composing music implies managing saliences, since it is these that bubble to the surface of a piece of music, participating in its character. A structural musical analysis examining

[4] A fourth category, *texture*, has not yet been developed.

[5] As an example of a pattern with motion, consider $ABABBABBBABBBBABBBBB$, where the B segments display a (patterned) growth.

the management of saliences has a chance of revealing something about how music may be composed and how it may be heard.

Simplicity provides a perceptual basis for music analysis; it also serves as an effective base case for reasoning, allowing a systematic analysis of basic musical material. While we cannot speak of maximally *complex* music, the *simplest* ways to make a piece of music can be enumerated.

Likewise, the simplest ways to make a pattern, a shape, or a motion can be enumerated. The simplest pattern is just repetition of a single term: $AAAAAA\ldots$. The simplest shapes (in a total ordering such as pitch, or loudness, or set cardinality) are those that are described by just one orientiation: *UP* or *DOWN* or *SAME*. Motion, as well, is simplest when it proceeds in just one direction.

1.2 Avoiding the Encoding Problem

The questions that follow are somewhat trickier: what is the *next simplest* pattern? what are the *next simplest* shapes? And what are the next simplest patterns and shapes after those?

It is tempting to devise a system to compose a few simple types such that (for example) *any* pattern can be encoded under the system.[6] After some initial exploration, we decided to avoid the "encoding problem," for the reasons that follow.

Under any coding system, there may be several ways of encoding a given pattern. We then must ask which encoding to prefer. A common solution is to prefer the shortest encoding, since it compresses the pattern as much as possible, thus making use of as much of the pattern's inherent structure as possible. However, the search for a shortest encoding is computationally hard (i.e. no efficient, deterministic algorithm is known).[7]

Even if we do manage to find (or approximate) a shortest encoding, we are left with a *measure* of pattern complexity (or entropy), and a single way of maximally compressing the pattern. Our problem, however, was not to *compress* the pattern, but to *describe* it in such a way that simplicities are indicated. This may turn out to take up *more* bits than the original pattern, since there may be many interesting things to point out which cannot all be summarized in a single encoding.

We may instead ask for a few different encodings, or all possible encodings. But this is still computationally hard, and still of limited utility for our problem. Imagine we have a pattern that looks like alphabet soup, but every time the term X appears, it is in a group of Xs such that each group is one term shorter than the last. Depending on our coding system and the algorithm we use for finding encodings, this regularity may or may not be expressed somewhere in our results. But even if it is, it will be located in some encoding of the entire pattern, and

[6] Something like this was attempted by [3], and much work was done along these lines by structural information theorists [4, 5].

[7] A non-deterministic, genetic approach to this problem is explored by [6].

we will still have the task ahead of us of searching through our encodings for simple patterns.

An alternative approach is to avoid the encoding of non-simple patterns, and instead to focus on finding simple structures within them. In the alphabet-soup example above, the system might not be able to provide a concise description of the entire pattern, but it should be able to point out that in the X dimension, at least, something simple is happening.

The total "dimensionality" of a piece of music is very large, and may be impossible to enumerate. A structural description therefore can't claim to be exhaustive. Instead, general methods are developed to unyoke structures into separate dimensions and to synchronize dimensions into higher-order dimensions. The separation and resynchronization may be directed by a higher-level program, randomized, guided by the simplicities discovered, or dictated by a combination of these methods. The composable type system allows any number of dimensions of pattern, shape, and motion to be explored without requiring the definition of new types to represent them, or of new functions for discovering and analyzing them.

In the remainder of this paper, we discuss a few primitive types and means of combination for musical shape, pattern, and motion. Because space is limited and work is ongoing, the description provided here is not of a complete system.

2 Shape

Shape can be built on the basis of *orientation*. Oriented shapes occur everywhere in every oriented (or quantifiable) aspect of music, including pitch, loudness, speed, density, and so on.

A simple shape called a *chain* is described by just one orientation, *UP*, *DOWN*, or *SAME*. Given a melody, it is easy to decompose its pitch content into a sequence of chains, with each chain overlapping the next by one pitch.

The *chain* is a simple type describing a local structure. It is recursively composable to describe large-scale structure. Recursive chains are called *Z-chains*.[8]

2.1 Recursive Orientation: Z-chains

First-order Z-chains (or Z-chains described by one orientation), are just chains. The first step in the composition of chains into second-order Z-chains is the unyoking of the three orientations into three dimensions. Thus, when chains are "chained together" to form higher-order chains, the principle of *chaining like to like* is observed.

Chains (and Z-chains) can be compared with respect to several different oriented features, including top pitch (more generally, top value), bottom pitch, chain length, and interval span. Z-chains are found in one feature at a time.

Having specified a feature and a sequence of nth order Z-chains in one (recursive) orientation, the same simple one-pass algorithm that finds chains is used

[8] The Z stands for the zig-zagging shapes that result.

to find Z-chains of the $(n+1)$th order. For example, given a sequence of chains going *UP* and comparing their top pitches, the chaining algorithm decomposes the sequence into Z-chains (in feature top pitch) with orientations *UP-UP*, *UP-DOWN*, and *UP-SAME*. The recursive Z-chain algorithm unyokes these new dimensions, and runs the chaining algorithm again on each.

Each pass of the chaining algorithm is shorter than the last, since at every subsequent level there fewer chains. The algorithm terminates when it finds only one chain of a given recursive orientation, since there remains nothing to chain it to. The algorithm for finding all Z-chains in a sequence (for a given feature) is efficient, taking time $O(n^2)$ where n is the number of items in the sequence.

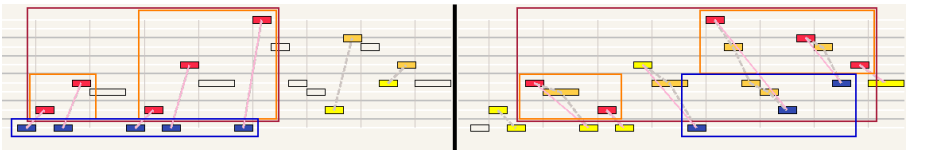


Fig. 1. Z-chains in "Happy Birthday." On the left, top pitches UP-UP-UP and bottom pitches UP-SAME; on the right, tops DOWN-DOWN-UP and bottoms DOWN-UP.

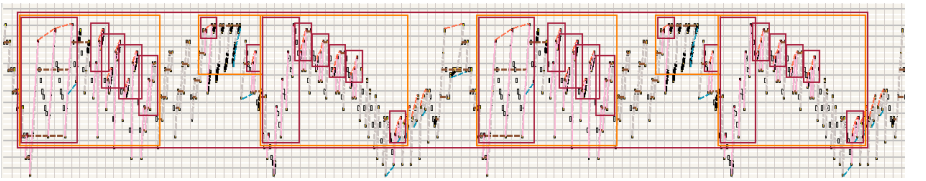


Fig. 2. A large scale fourth-order Z-chain in the first part of *Presto* from J. S. Bach, solo violin Sonata I. The outer box encapsulates the repetition.

Figures 1 and 2 illustrate the Z-chain concept over small and large scales, respectively.

Higher-order Z-chains may skip items in the sequence – in particular they skip Z-chains in orientations other than their own. This property allows us to infer that the orientation in question *does not exist* in a skipped stretch of music. For a long piece of music, very high-order Z-chains may provide an overview of some aspects of form, but they may also be fragmentary. Z-chains of second and third orders, on the other hand, tend to provide compact synopses of short stretches of music.

Z-chains are based on the concept of chaining *like to like*. A common motivation in computational music analysis is to search for *repetition*. This is often done by comparing over note n-grams, possibly with a tolerance for error to admit variation. The Z-chain scheme of chaining like to like, on the other hand, is an

efficient search for *parallelism*, allowing similar structures (for a strictly-defined definition of "similar", *not* an error tolerance) to be linked. The link is itself *oriented*, so that instead of saying of a structure "here it is, and here it is again" (as in the search for repetition), a higher-level structure is constructed in which the component lower-level structures stand in a specified oriented relation to one another.

2.2 Synchronizing Z-chains

A synchronization function is a means of combining structures in different (orthogonal) dimensions. *Mini-schemas* are made from inclusions (overlaps) of Z-chains in different orientable features (i.e. top pitch, bottom pitch, chain length, etc.). Mini-schemas are specified by Z-chains in at least two different features, and instantiated (as Z-chains are) as lists of chains.

Mini-schemas are constrained to consist of an *entire* Z-chain of some order in (at least) one feature, and to consist of *consecutive* first-order chains. Given a feature-set minimally including "top pitch" and "bottom pitch," the union of all mini-schemas always covers the entire sequence, since every two consecutive pitch chains have oriented top pitches and bottom pitches.

A mini-schema specification may be instantiated more than once in a given piece. The multiple instances of a schema need not be note-for-note identical, they are only "the same" with respect to their multi-feature Z description (and their compactness with regards to chains). They need not be the same in all features, but only in the subset of features which is used in their description. They need not contain the same number of first-order chains.

If there is more than one instance of a given mini-schema, it is possible to compare instances to discover how the schema is deployed throughout the course of the piece.

Mini-schemas are illustrated in Figure 3.

3 Pattern

A fact of music is that difference is valued: composers are obliged to produce music that does not repeat *any* known music. It is also true that within a single piece, differentiation is important, since it is difference that articulates form (at all scales) and through which discourses occur.

In contrast with oriented shape, in which any two terms are comparable within a total ordering, a comparison of *pattern* terms tells us only whether they are equal or unequal. In a pattern like ABA, nothing is known about the relationship between A and B, except that they are different, while A and A are the same.

An obvious mode of pattern analysis is the detection of recurrent subsequences, for which well-known methods apply. A second mode involves the recognition of "concentric" patterns, in which the ABA pattern is generalized by considering B as an *island* in a sea of A. Similarly, the following pattern has an island of C: *(ABBA(CC)ABABAA)*.

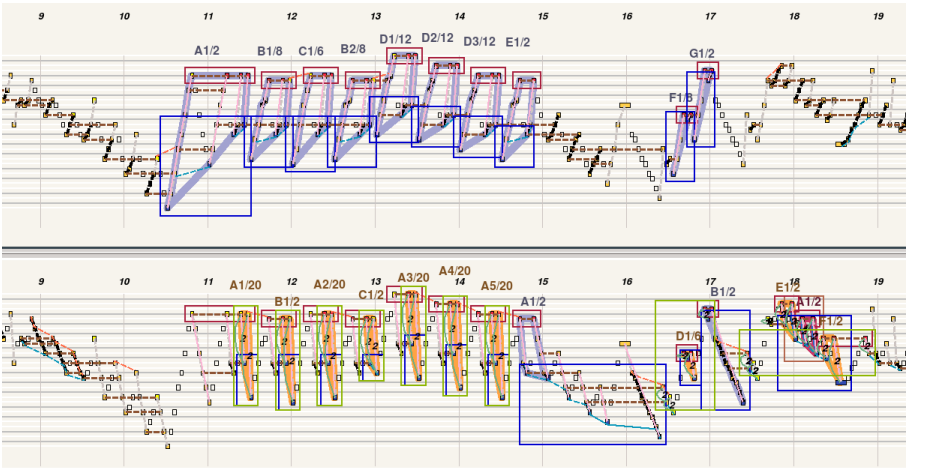


Fig. 3. Schemas in J. S. Bach: *Gigue*, Solo violin Partita II. Top half is view of chains UP, bottom half is chains DOWN. The schemas outlined in lilac in the top half are defined by constant pitch top, rising pitch bottom. The orange schemas in the lower view are "hotter", comprised of the additional feature of constant chain length, shown with green arcs.

The concentric-pattern algorithm has, as a natural consequence, the effect of segmenting stand-alone sections such as *(ABCABCABCABC)(DEFDEF)*, in which no terms in any top level group occur in any other top level group.

It is possible to generate patterns from schemas by comparing instances of a schema specification under an equality predicate (e.g. exact equality, equality under transposition, etc.). Among these are sure to be some easily identified as "canonical," fully interpretable as structures of repetition within concentricities. The segmentation of a Bach work composed of two repeating sections proves to be a trivial consequence of this analysis.

4 Motion

Motion naturally leads us to inquire into rhythm, but here we restrict ourselves to the motion of schemas. Schemas allow for the construction of "motion shapes" in the following way. Schemas are partial descriptions: nothing precludes a description involving three features partially covering a description with just two features of the three. We may presume that the schema with more synchronized features is more regular than one with fewer: in McLuhan's (jazz-influenced) terminology, the more regular schema is the "hotter" pattern, where a "hot pattern" is said to "drive" perception and a "cool" pattern is said to invite perceptual completion. The partial covering of a "cool" pattern – involving few features – by a "hotter" pattern – a superset of the cool pattern – can be thought of as a heat transition. Figure 4 shows an example.

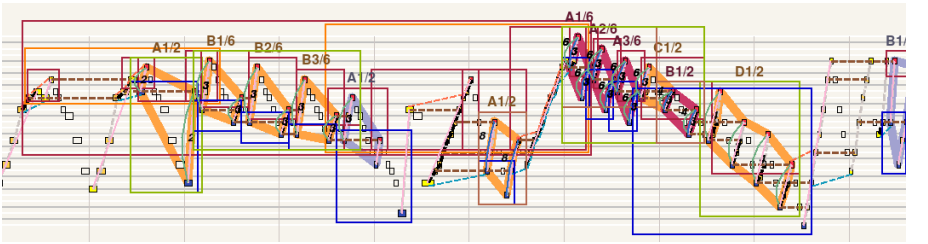


Fig. 4. *Gigue*, near the opening. Differential heat in the same schema, comprising *heat transitions*. Red is hottest, i.e most regular, followed by orange and lilac. The boxes show the Z-chains synchronized by the schemas.

5 Reasoning About Music

One benefit of the foregoing system lies in its ability to bring out formal structure in arbitrarily simple or complex music, constructing divisions based on pattern-recurrence of schemas. A further goal is a much more extensive elaboration showing the interrelation of the parts of music.

The aim, in the material presented, was the generation of a "primary" (and deterministic) level of musical objects. More objects are constructed by discovering further relations, generating more shapes and patterns, at which point the system is fully re-entrant. For example, each instance of a schema yields a pattern of *length* in its base chains: the pattern can be treated as a Z-chain, whose synchronicity with any other features can be schematized. From this point on, analysis may proceed *non-deterministically*, since object generation is potentially unlimited, and the possibilities of synchronization are vast.

References

1. Marr, D.: Vision: A computational investigation into the human representation and processing of visual information. W. H. Freeman, San Francisco (1982)
2. Cohen, H.: The further exploits of Aaron, painter. Stanford Humanities Review, 4.2 (1995)
3. Simon, H.A., Sumner, R.K.: Patterns in music. In: B. Kleinmuntz (Ed.) Formal representation of human judgement. Wiley, New York (1968)
4. Leeuwenberg, E.L.J.: A perceptual coding language for visual and auditory patterns. American Journal of Psychology, 84, 307-349 (1971)
5. van der Helm, P.A.: Cognitive architecture of perceptual organization: From neurons to gnosons. Cognitive Processing, 13, 13-40 (2012)
6. Dastani, M., Marchiori, E., Voorn, R.: Finding Perceived Pattern Structures using Genetic Programming. In: Genetic and Evolutionary Computation Conference (2001)

On Automatic Music Genre Recognition by Sparse Representation Classification using Auditory Temporal Modulations

Bob L. Sturm[1] and Pardis Noorzad[2]

[1] Department of Architecture, Design and Media Technology
Aalborg University Copenhagen
Lautrupvang 15, 2750 Ballerup, Denmark
[2] Department of Computer Engineering and IT
Amirkabir University of Technology
424 Hafez Ave., Tehran, Iran
bst@create.aau.dk, pardis@aut.ac.ir

Abstract. A recent system combining sparse representation classification (SRC) and a perceptually-based acoustic feature (ATM) [31, 30, 29], is reported to outperform by a significant margin the state of the art in music genre recognition, e.g., [3]. With genre so difficult to define, this remarkable result motivates investigation into, among other things, why it works and what it means for how humans organize music. In this paper, we review the application of SRC and ATM to recognizing genre, and attempt to reproduce the results of [31] where they report 91% accuracy for a 10-class dataset. We find that only when we pose the sparse representation problem with inequality constraints, and, more significantly, reduce the number of classes by half, do we begin see accuracies near those reported. In addition, we find evidence that this approach to classification does not benefit significantly from the features being based on a perceptual analysis.

1 Introduction

Simply because we lack clearly definitive examples, and any utilitarian definitions, the automatic recognition of music genre is different from other tasks in music information retrieval. The human categorization of music seems natural, yet appears fluid and often arbitrary by the way it appears motivated by more than measurable characteristics of audible changes in pressure [9, 26, 17]. Extra-musical information, such as artist fashion, rivalries and the fan-base, associated dance styles, lyrical subjects, societal and political factors, religious beliefs, and origins in time and location, can position a particular piece of music into one category or another, not often without debate [38]. With the changing fashions of communities and the needs of companies, new genres are born [17]. And genres become irrelevant and lost, though we might still hear the recorded music and classify it as something entirely different.

It seems daunting then to make a computer recognize genre with any success. Yet, in developments between 2002 and 2006 [23], we have seen the accuracy of such algorithms progress from about 60% — using parametric models created from bags of features [43] — to above 80% — aggregating features over long time scales and boosting weak classifiers [3]. The majority of approaches developed

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

so far use features derived only from the waveform, and/or its symbolic form. Some work has also explored mining user tags [27] and written reviews [1], or analyzing song lyrics [22]. Since humans have been measured to have accuracies around 70% after listening to 3 seconds of music — which surprisingly drops only down to about 60% for only half a second of listening [13] — the results of the past decade show that the human categorization of music appears grounded to a large extent in acoustic features, at least at some coarse granularity.

Recently, we have seen a large leap in genre classification accuracy. In [31, 30, 29], the authors claim that with a perceptually-motivated acoustic feature, and a framework of sparse representation classification (SRC) [46], we move from 82.5% accuracy [3] to up to 93.7%. SRC, which has produced very promising results in computer vision [46, 47] and speech recognition [11, 35], can be thought of as a generalization of k-nearest neighbors (kNN) for multiclass classification with many important advantages. It is a global method, in the sense that it classifies based on the entire training set; and it does not rely only on local similarity information as does kNN. SRC can prevent overcounting of neighborhood information by virtue of its emphasis on sparsity in the representation. Additionally, SRC assigns a weight to each training set sample, thus quantifying the degree of its importance. All of these points make SRC a strong classification method.

The massive improvement in genre recognition that accompanies this approach motivates many questions, not only about what is working and why it is working so well, but also about how we perceive rich acoustic scenes, and the way we think and talk about music. For instance, are these purely acoustic features so discriminative because they are modeled on the auditory system of humans? Since the features in [31] are computed from segments of very long duration (30 s), how robust is the method to shorter observations, e.g., can it reach 60% for 500 ms? Do its misclassifications make sense, and to some extent forgivable? Do the features cluster in a sensible way, and do subgenres appear as smaller clusters within larger clusters? Can we compute high-level descriptors from these features, such as rhythm, harmony, or tempo?

In this work, we review the approach proposed in [31], and describe our attempt to reproduce the results, making explicit the many decisions we have had to make to produce the features, and to build the classifier. The accuracies we observe, however, are 30–40% inferior to those in [31], even with the improvement we observe when posing the sparse representation problem using inequality constraints rather than the equality constraints specified in [31, 30, 29]. Only when we reduce by half the number of classes tested in [31] do we see the reported high accuracies. In addition, we find evidence that the perceptual nature of the features has no significant impact on the classifier accuracy. We make available our MATLAB code, both classification and feature extraction, with which all results and figures in this article can be reproduced: `http://imi.aau.dk/~bst/software/`.

2 Background

We now review SRC from a general perspective, and then we review modulation analysis for feature extraction, and its application specifically to music genre

recognition. Throughout, we work in a real Hilbert space with inner product $\langle \mathbf{x}, \mathbf{y} \rangle := \mathbf{y}^T \mathbf{x}$, and p-norm $\|\mathbf{x}\|_p^p := \sum_i |[\mathbf{x}]_i|^p$, for $p \geq 1$, where $[\mathbf{x}]_i$ is the ith component of the column vector $\mathbf{x}$.

2.1 Classification via sparse representation in labeled features

Define a set of N labeled features, each belonging to one of C enumerated classes

$$\mathcal{D} := \{(\mathbf{x}_n, c_n) : \mathbf{x}_n \in \mathbb{R}^m, c_n \in \{1, \ldots, C\}\}_{n \in \{1,\ldots,N\}}. \tag{1}$$

And define $\mathcal{I}_c \subset \{1, \ldots, N\}$ as the indices of the features in $\mathcal{D}$ that belong to class c. Given an unlabeled feature $\mathbf{y} \in \mathbb{R}^m$, we want to determine its class using $\mathcal{D}$. In kNN, we assume that the neighborhood of $\mathbf{y}$ carries class information, and so we classify it by a majority vote of its k-nearest neighbors in $\mathcal{D}$. Instead of iteratively seeking the best reconstruction of $\mathbf{y}$ by a single training sample (i.e., its ith nearest neighbor), we find a reconstruction of $\mathbf{y}$ by all training samples. Then we choose the class whose samples contribute the most to the reconstruction. We have SRC when we enforce a sparse reconstruction of $\mathbf{y}$.

SRC essentially entails finding nearest to an unlabeled feature its linear approximation by class-restricted features. To classify an unlabeled feature $\mathbf{y}$, we first find the linear combination of features in $\mathcal{D}$ that constructs $\mathbf{y}$ with the fewest number of non-zero weights, regardless of class membership, posed as

$$\min_{\mathbf{a} \in \mathbb{R}^N} \|\mathbf{a}\|_0 \;\text{ subject to }\; \mathbf{y} = \mathbf{D}\mathbf{a} \tag{2}$$

where we define the $m \times N$ matrix $\mathbf{D} := [\mathbf{x}_1 | \mathbf{x}_2 | \ldots | \mathbf{x}_N]$, and the pseudonorm $\|\mathbf{a}\|_0$ is defined as the number of non-zero weights in $\mathbf{a} := [a_1, a_2, \ldots, a_N]^T$. We might not want to enforce equality constraints, and so we can instead pose this

$$\min_{\mathbf{a} \in \mathbb{R}^N} \|\mathbf{a}\|_0 \;\text{ subject to }\; \|\mathbf{y} - \mathbf{D}\mathbf{a}\|_2^2 \leq \epsilon^2 \tag{3}$$

where $\epsilon^2 > 0$ is a maximum allowed error in the approximation. All of this, of course, assumes that we are using features that are additive. We can extend this to non-linear combinations of features by adding such combinations to $\mathcal{D}$ [35], which can substantially increase the size of the dictionary.

We now define the set of class-restricted weights $\{\mathbf{a}_c\}_{c \in \{1,2,\ldots,C\}}$

$$[\mathbf{a}_c]_n := \begin{cases} a_n, & n \in \mathcal{I}_c \\ 0, & \text{else.} \end{cases} \tag{4}$$

The non-zero weights in $\mathbf{a}_c$ are thus only those specific to class c. From these, we construct the set of C approximations and their labels $\mathcal{Y}(\mathbf{a}) := \{\hat{\mathbf{y}}_c(\mathbf{a}) := \mathbf{D}\mathbf{a}_c\}_{c \in \{1,2,\ldots,C\}}$, and we assign a label to $\mathbf{y}$ simply by a nearest neighbor criterion

$$\hat{c} := \arg \min_{c \in \{1,\ldots,C\}} \|\mathbf{y} - \hat{\mathbf{y}}_c(\mathbf{a})\|_2^2. \tag{5}$$

Thus, SRC picks the class of the nearest approximation of $\mathbf{y}$ in $\mathcal{Y}(\mathbf{a})$.

We cannot, in general, efficiently solve the sparse approximation problems above [8], but there exist several strategies to solve them. We briefly review the convex optimization approaches, but [42] provides a good overview of many more; and [48] is a large study of SRC using many approaches. Basis pursuit (BP) [6] proposes relaxing strict sparsity with the convex ℓ_1-norm

$$\min_{\mathbf{a}\in\mathbb{R}^N} \|\mathbf{a}\|_1 \;\text{ subject to }\; \mathbf{y} = \mathbf{D}\mathbf{a}. \tag{6}$$

And without equality constraints, BP denoising (BPDN) [6] poses this as

$$\min_{\mathbf{a}\in\mathbb{R}^N} \|\mathbf{a}\|_1 \;\text{ subject to }\; \|\mathbf{y} - \mathbf{D}\mathbf{a}\|_2^2 \leq \epsilon^2. \tag{7}$$

One could also change the ℓ_2 error to ℓ_1 to promote sparsity in the error [47, 12]. We have the LASSO [41] when we switch the objective and constraint of BPDN

$$\min_{\mathbf{a}\in\mathbb{R}^N} \|\mathbf{y} - \mathbf{D}\mathbf{a}\|_2^2 \;\text{ subject to }\; \|\mathbf{a}\|_1 \leq \rho \tag{8}$$

where $\rho > 0$. Furthermore, we can pose the problem in a joint fashion

$$\min_{\mathbf{a}\in\mathbb{R}^N} \frac{1}{2}\|\mathbf{y} - \mathbf{D}\mathbf{a}\|_2^2 + \lambda\|\mathbf{a}\|_1 \tag{9}$$

where $\lambda > 0$ tunes our preference for sparse solutions versus small error.

Along with using the ℓ_1 norm, we can reduce the dimensionality of the problem in the feature space [47]. For instance, the BPDN principle (7) becomes

$$\min_{\mathbf{a}\in\mathbb{R}^N} \|\mathbf{a}\|_1 \;\text{ subject to }\; \|\mathbf{\Phi}\mathbf{y} - \mathbf{\Phi}\mathbf{D}\mathbf{a}\|_2^2 \leq \epsilon^2 \tag{10}$$

where $\mathbf{\Phi}$ is a fat full-rank matrix mapping the features into some subspace. To design $\mathbf{\Phi}$ such that the mapping might benefit classification, we can compute it using information from $\mathbf{D}$, e.g., by principal component analysis (PCA) or non-negative matrix factorization (NMF), or we can compute it non-adaptively by random projection. With PCA, we obtain an orthonormal basis describing the directions of variation in the features, from which we define $\mathbf{\Phi}^T$ as the $d \leq m$ significant directions, i.e., those having the d largest principal components.

Given $d \leq m$, NMF finds a positive full rank $m \times d$ matrix $\mathbf{U}$ such that

$$\min_{\mathbf{U}\in\mathbb{R}_+^{m\times d}} \frac{1}{N}\sum_{n=1}^{N} \|\mathbf{x}_n - \mathbf{U}\mathbf{v}_n\|_2^2 \text{ subject to } \mathbf{v}_n \succeq 0. \tag{11}$$

The full-rank matrix $\mathbf{U}$ contains d templates that approximate each feature in $\mathbf{D}$ by an additive combination. Thus the range space of $\mathbf{U}$ provides a good approximation of the features in $\mathbf{D}$, with respect to the mean ℓ_2-norm of their errors. In this case, we make $\mathbf{\Phi}^T := (\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T$.

Finally, we can reduce feature dimensionality by random projection [7, 4, 21], where we form the entries of $\mathbf{\Phi}$ by sampling from a random variable, e.g., Normal, and without regard to $\mathbf{D}$. We normalize the columns to have unit ℓ_2-norm, and ensure $\mathbf{\Phi}$ has full rank. While this approach is computationally simple, its non-adaptivity can hurt classifier performance [21].

2.2 Modulation Analysis

Modulation representations of acoustic signals describe the variation of spectral power in scale, rate, time and frequency. This approach has been motivated by the human auditory and visual systems [44, 15, 36, 40, 24]. In the literature, we find two types of modulation representations of acoustic signals, which seemingly have been developed independently. One might see these approaches as a form of feature integration, which aggregate a collection of small scale features.

In [44, 24], the authors model the output of the human primary auditory system as a multiscale spectro-temporal modulation analysis, which [44] terms a "reduced cortical representation" (RCR). To generate an RCR, one first produces an "auditory spectrogram" (AS) approximating the time-frequency distribution of power at the output of the early stage of the auditory system [49]. This involves filtering the signal with bandpass filters modeling the frequency responses of the hair cells along the basilar membrane, then calculating activations of the nerve cells in each band, and finally extracting a spectral power estimate from the activation patterns [49, 44, 24]. In the next step, which models the central auditory system, one performs a "ripple analysis" of the AS, giving the local magnitudes and phases of modulations in scale and modulation rate over time and frequency [44, 24]. This procedure uses 2-D time-frequency modulation-selective filters, equivalent to a multiresolution affine wavelet analysis sensitive to fast and slow upward and downward changes in frequency [44]. To obtain spectro-temporal modulation content [24], one integrates this four-dimensional representation over time and/or frequency.

A similar representation is proposed in [15], where the authors extract modulation information by applying a Fourier transform to the output of a set of bandpass filters modeling the basilar membrane. The magnitude output of this gives a time-varying modulation spectrogram. One could instead apply a wavelet transform to each row of a magnitude spectrogram, and then integrate the power at each scale of each band along the time axis. This produces a modulation rate-scale representation [40].

Motivated by its perceptual foundation [49, 44], and success in automatic sound discrimination [45, 24], the work of [28] appears to be the first to use modulation analysis features for music genre recognition, which they further refine in [31, 30, 32]. In [28], the authors use the toolbox by the Neural Systems Laboratory (NSL) [33, 34] to derive an RCR and perform a ripple analysis. They then average this across time to produce tensors of power distributed in modulation rate, scale, and acoustic frequency. While the features in that work are built from the RCR [44, 24], the features used in [31, 30] are a joint scale-frequency analysis [40] of an AS created from the model in [49]. This feature, which they call an "auditory temporal modulation" (ATM), describes power variation over modulation scale in each primary auditory cortex channel.

3 Recreating the Features and Classifier of [31]

In this section, we first describe how we generate ATM features, which are described in part in [31, 32]; and then we describe the approach to classify an ATM using SRC presented in part in [31].

3.1 Building Auditory Temporal Modulations

The authors take 30 seconds of music, downsample it to 16 kHz, then make it zero mean and unit variance. They then compute an AS following the model of the primary auditory system of [49], except they use a constant-Q transform of 96 bandpass filters covering a 4-octave range (24 filters per octave), whereas [49] uses an affine wavelet transform of 64 scales covering 5 octaves from about 173 Hz to 5.9 kHz. Finally, they pass each channel of the AS through a Gabor filterbank sensitive to particular modulation rates, and form the ATM by integrating the energy output at each filter.

To create ATMs, we have tried to follow as closely as possible the description in [31, 32]. We first generate a constant-Q filter bank with 97 bands spaced over a little more than four octaves, with $N_f = 24$ filters per octave. We center the first filter at 200 Hz because that is specified in [49]. The last filter is thus centered on 3200 Hz. Since in [49] the model of the final stage of the primary audio cortex computes first-order derivates across adjacent frequency bands, we end up with a 96 band AS as specified in [31, 32].

We create our constant-Q filter bank as a set of finite impulse response filters designed by the windowing method [25]. Since it is not mentioned in [49, 31, 32], we make all filters independent, and to have the same gain. To generate the impulse responses of our filterbank, we modulate a prototype lowpass window to logarithmically spaced frequencies. Because of its good low passband characteristic, we use a Hamming window, which for the kth filter ($k \geq 1$) produces the impulse response sampled at F_s Hz

$$h_k(n) := \gamma_k \left[0.54 - 0.46 \cos\left(\frac{2\pi n}{l_k}\right)\right] e^{j2\pi\omega_k n/F_s},\ 0 \leq n < l_k \tag{12}$$

with a modulation frequency $\omega_k := f_{\min} 2^{(k-1)/N_f}$ Hz, and length in samples

$$l_k := \left\lceil \frac{q}{2^{k/N_f} - 2^{(k-1)/N_f}} \frac{F_s}{f_{\min}} \right\rceil. \tag{13}$$

We set the gain γ_k such that there is no attenuation at the kth center frequency, i.e., $|\mathcal{F}\{h_k(n)\}(\omega_k)| = 2$, where $\mathcal{F}\{x(n)\}(\omega)$ is the Fourier transform of $x(n)$ evaluated at frequency ω. The factor $q > 0$ tunes the width of the main lobe. We choose $q \approx 1.316$ such that adjacent filters overlap at their -3 dB stopband.

This model of the basilar membrane is simplified considering its non-adaptive and uniform nature, e.g., it does not take into account masking and equal loudness curves. An alternative model of the cochlea is given by Lyon [20], which involves a filterbank with center frequencies spread uniformly below a certain frequency, and logarithmically above [37]. Figure 1 shows that the Lyon model attenuates single sinusoids at frequencies tuned to the center frequencies of its filterbank. Our filterbank uniformly passes these frequencies, albeit over a smaller four octave range [31, 32] assumed to begin at 200 Hz. Figure 1 also shows that the filterbank of the NSL model [34] by and large has a uniform attenuation.

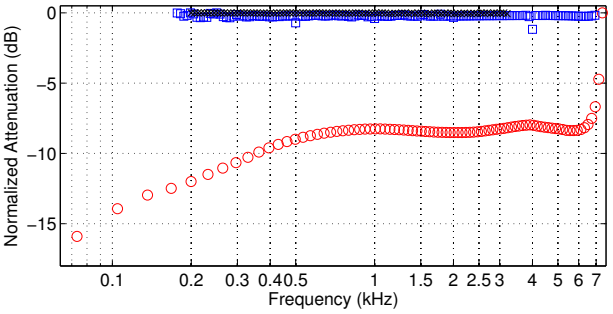


Fig. 1. Attenuations of single sinusoids with the same power, at frequencies identical to center frequencies in the filterbanks. (x) Our constant-Q filter bank. (o) Lyon passive ear model [20, 37]. (□) NSL ear model [34].

We pass though our constant-Q filter bank a sampled, zero-mean and unit-variance acoustic signal $y(n)$ [31, 32], which produces for the kth filter the output

$$y_k(n) := \sum_{m=0}^{l_k-1} h_k(m) y(n - m - \Delta_k) \tag{14}$$

where $\Delta_k > 0$ is the group delay of the kth filter at ω_k. This delay correction is necessary because the filters we use to model the basilar membrane have different lengths. This correction is unnecessary in the implementation of the Lyon [37] or NSL models [34], since they use second-order sections with identical delays.

As in [31, 32], we next take the sample wise difference in each band

$$y'_k(n) := y_k(n) - y_k(n-1). \tag{15}$$

which models the action potential of the hair cell [49]. This now goes through a non-linear compression, followed by a low pass filter modeling leakage in the hair cell membrane. Referring to [49], we see the compression can be modeled as a sigmoidal function, and that the output of the kth channel is

$$g_k(n) := \frac{1}{1 + e^{-\gamma y'_k(n)}} - \frac{1}{2} \tag{16}$$

where $\gamma > 0$ is depends on sound pressure level [49]. Furthermore, "... saturation in a given fiber is limited to 30-40 dB" [49], implying γ is somehow set adaptively. In reality, we cannot equate the values of the digital samples in $y'_k(n)$ with the physical pressure embodied in this compression. However, working naively, we might absorb into γ such a conversion, and find some value that actually compresses. Figure 2 shows the cumulative distribution of amplitudes input to the compressor (15) with a 30 second music signal having unit energy [31, 32]. For $\gamma = 1$, we see that this distribution is compressed, whereas setting $\gamma = 10$ results in an expansion. Thus, we set $\gamma = 1$ independent of the input, and assume it compresses $y'_k(n)$ from any 30 second music signal scaled to have unit energy.

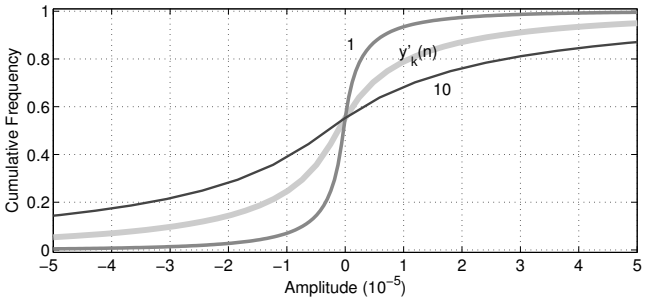


Fig. 2. Cumulative distributions of amplitude input to compressor ($y'_k(n)$), and output as a function of γ (labeled).

The compressor output $g_k(n)$ is then smoothed by the hair cell membrane and attendant leakage [49, 32], which passes frequencies only up to 4 – 5 kHz [49]. Thus, we pass each $g_k(n)$ through a 6th-order Butterworth filter having a cutoff frequency of 4 kHz, producing $f_k(n)$. This is then processed by a "lateral inhibitory network," described in [49], which detects discontinuities in the response. This entails a spatial derivative across channels with smoothing, a half-wave rectifier, and then integration; but [31, 32] does not specify smoothing, and states the process can be approximated by a first order derivative across logarithmic frequency. Thus, we compute for channel $s \in \{1, \ldots, 96\}$

$$v_s(n) := [f_{s+1}(n) - f_s(n)]\mu[f_{s+1}(n) - f_s(n)] \tag{17}$$

where $\mu(u) = 1$ if $u \geq 0$, and zero otherwise.

In the final step, we integrate the output with "a [possibly rectangular window with a] long time constant (10-20 ms)" [49], or a 2 – 8 ms exponential window [31, 32]. Thus, we compute the nth sample of the kth row of the AS by

$$A_k(n) := \sum_{m=0}^{\lfloor F_s \tau \rfloor} v_s(n-m) e^{-m/F_s\tau} \tag{18}$$

where we define $\tau := 8$ ms. This completes the first step of building an ATM.

Figure 3 compares the resulting AS from our model built from interpreting [49, 31, 32], that of the auditory model designed by Lyon [20, 37], and the cortical representation from the NSL model [33, 34]. The Lyon model uses 86 bands non-uniformly spread over a little more than 6.5 octaves in 80 – 7630 Hz [20, 37], whereas the NSL model covers 5.33 octaves with 24 filters per octave logarithmically spread over 180 – 7246 Hz [33, 34]. Though the frequency range of those models are larger, we only use a 4-octave range as in [31, 32].

To generate an ATM, [31, 32] describe first performing a multiresolution wavelet decomposition of each row of an AS, and then integrating the squared output across the translation axis. Based on experimental evidence [36], the authors use a set of Gabor filters sensitive to eight modulation rates $\{2, 4, 8, \ldots, 256\}$

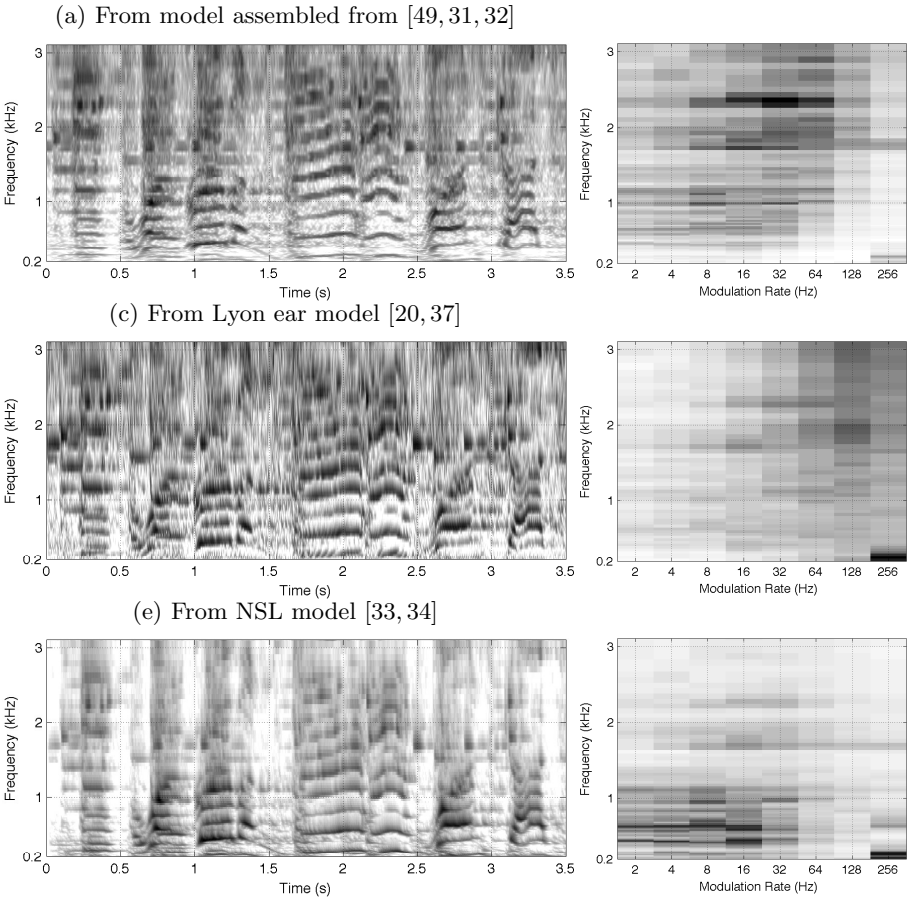


Fig. 3. Auditory spectrograms (left) and their auditory temporal modulations (right).

Hz. We assume this Gabor filterbank can be assembled as follows. We define the sampled impulse response truncated to length N_l of our complex Gabor filter tuned to a modulation rate $f_0 2^l \geq 0$ Hz, and of scale $F_s\alpha/f_0 2^l > 0$

$$\psi(n; f_0 2^l) := \frac{f_0 2^l}{F_s \alpha} \left[e^{-(f_0 2^l/\alpha)^2((n-N_l/2)/F_s)^2} e^{j2\pi f_0 2^l n/F_s} - \mu_l \right] \tag{19}$$

for $n = 0, \ldots, N_l - 1$, where we define μ_l such that $\psi(n; f_0 2^l)$ has zero mean. The normalization constant assures uniform attenuation at each modulation frequency, as used in joint scale-frequency analysis [40]. We set $\alpha = 256/400$ and $N_l = 4F_s\alpha/f_0 2^l$. Since a Gabor filter tuned to a low frequency has a high DC component, we make each row of the AS zero mean, thus producing $A'_k(n)$. Passing the kth row of this AS through the lth channel ($l \in \{0, 1, \ldots, 7\}$) of the Gabor filterbank produces the convolution $R_{k,l}(n) := [\psi(m; f_0 2^l) \star A'_k(m)](n)$. Finally, as in [31, 32], we sum the squared modulus of the output sampled at all

wavelet translations, producing the (k, l) element of the ATM

$$[\mathbf{A}]_{kl} := \sum_{p \in \mathbb{Z}} |R_{k,l}(p \lfloor F_s \alpha / f_0 2^{l+1} \rfloor)|^2 \tag{20}$$

where p is an integer multiplying the wavelet translations, which we assume is half the wavelet scale.

To the right of each AS in Fig. 3 we see the resulting ATM. Portions of these ATMs appear similar, with major differences in scaling and feature dimensionality. Within the four octave range specified in [31, 32], the dimensionality of the vectorized features are: 768 for the ATM in [31, 32], 416 for that created from the model by Lyon [20, 37], and 800 using the NSL model [33, 34].

3.2 Classifying Genre by Auditory Temporal Modulations

Given a set $\mathcal{D}$ of vectorized ATM features, each associated with a single music genre, we can use the machinery of SRC to label an unknown vectorized ATM $\mathbf{y}$. Following [31], we first make all features of $\mathcal{D}$ have unit ℓ_2-norm, as well as the test feature $\mathbf{y}$. We next solve the BP optimization problem posed in [31]

$$\min_{\mathbf{a} \in \mathbb{R}^N} \|\mathbf{a}\|_1 \;\; \text{subject to} \;\; \mathbf{\Phi y} = \mathbf{\Phi D a} \tag{21}$$

where $\mathbf{\Phi}$ reduces the features by, e.g., PCA. Finally, to classify $\mathbf{y}$, we construct the set of weights in (4), and assign a single genre label using the criterion (5).

Since we are working with real vectors, we can solve (21) as a linear program [6], for which numerous solvers have been implemented, e.g., [5, 10, 2, 14]. Because of its speed, we choose as the first step the root-finding method of the SPGL1 solver [2]. If this fails to find a solution, then we use the primal-dual method of ℓ_1-Magic [5], which takes as its starting point the minimum ℓ_2-norm solution $\mathbf{a}_2 := (\mathbf{\Phi D})^\dagger \mathbf{y}$. This initial solution satisfies the constraints of (21) as long as $\mathbf{\Phi D}$ has full rank, but probably is not the optimal solution. If the solution $\hat{\mathbf{a}}$ does not satisfy $\|\mathbf{\Phi y} - \mathbf{\Phi D}\hat{\mathbf{a}}\|_2^2 < 10^{-16}$ (numerical precision), we set $\hat{\mathbf{a}} := \mathbf{a}_2$.

4 Experimental Results

As in [31], we use the music genre dataset of [43] (GTZAN),[3] which has 1000 half-minute sound examples drawn from music in 10 broad genres: blues, classical, country, disco, hiphop, jazz, metal, pop, reggae, and rock. We define $\mathbf{\Phi}$ by PCA, NMF, or random sampling; and as in [31], we test dimension reduction by factors of $\{64, 16, 8, 4\}$, e.g., we reduce a feature vector of 768 dimensions by a factor of four to 192 dimensions. We also test downsampling the features, but we define it as vectorizing the result of lowpass filtering and decimating each column of the ATM (20). It is not clear how downsampling is done in [31]. In our case, a factor of f downsampling results in a vectorized feature of dimension $8\lceil 96/f \rceil$ when using our 96-channel features. Finally, as done in [3, 31], we use stratified 10-fold cross-validation for classifier training and testing.

[3] Available at: `http://marsyas.info/download/data_sets`

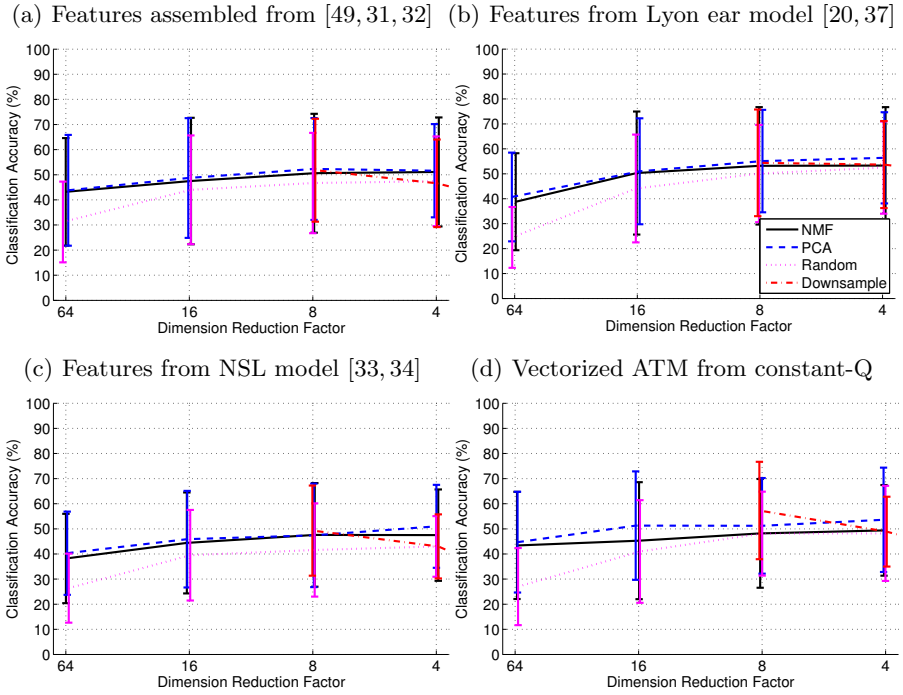


Fig. 4. Mean classification accuracy (10 classes) of SRC based on (21) for four different feature design methods, four dimension reduction methods, and several reduction factors. Overlaid is the standard deviation. (We add a slight x-offset to each bar for readability.)

Figure 4 shows our classification results for four different features, including the vectorized modulation-analysis of the magnitude output of the constant-Q filterbank that precedes (15). Across all features and dimension reduction methods and factors, we see no mean accuracies above 57.3% — which is produced by using features that do not model the entire primary auditory cortex. Since we see all mean accuracies are within one standard deviation of each other, we cannot claim one feature, reduction method is performing significantly different from any other. This result has been observed before in the application of SRC to face recognition [47]. In the experimental results of [31], however, we see features reduced a factor of 4 by NMF give the best results: mean accuracy of around 91% with a standard deviation of 1.76%. From the plots in [31], we can surmise there to be a statistically significant (e.g., $\alpha := 0.05$) difference between the features reduction methods. This contradicts our results and those of [47], not to mention the significant difference between the best accuracies on this same dataset with the same experimental protocol.

We have verified every part of our system is working as expected. We have performed modulation analysis on synthetic signals with known modulations. We have tested and confirmed on a handwritten digits dataset [16] that the SRC classifier performs comparably to other classifiers, and that our feature reduction is working. In this context too, we find no signficant difference in performance

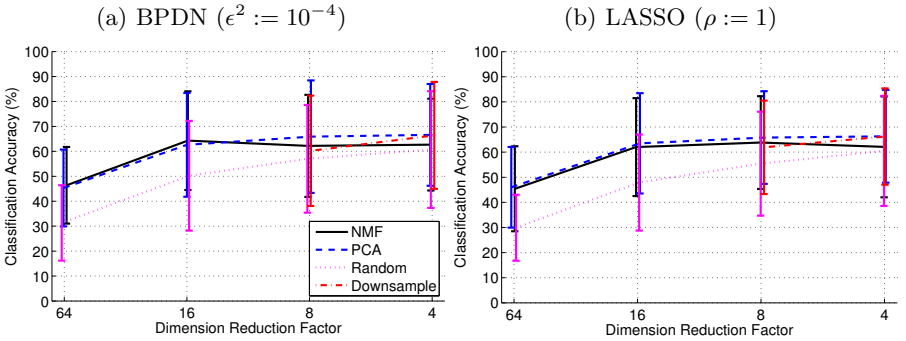


Fig. 5. Mean classification accuracy (10 classes) of SRC based on the BPDN (22) and LASSO (23), with features having dimensions mapped to $[0, 1]$, and a normalized projected dictionary, for ATM features from the Lyon model [20, 37], four dimension reduction methods, and several reduction factors. Overlaid is the standard deviation.

between feature reduction methods. From our experimentation, and conversation with the authors of [31], we believe that these differences come from several things, three of which are significant.

First, it is common in machine learning to preprocess features by accounting for dimensions with different scales. Panagakis et al. state that they make the values of each row of $\mathbf{D}$ be in $[0, 1]$ by finding and subtracting the minimum, and then dividing by the difference of the maximum and minimum.[4] When we rerun the experiments above with this modified data, we see the mean accuracy increases, but does not exceed the highest of 64% for the NSL features reduced in dimensionality a factor of 4 by PCA. Again, we see no significant difference between classifier performance with these features.

The second problem is posing the sparse representation with equality constraints in (21), which forces the sparse representation algorithm to model a feature exactly when instead we just want to find a good model of our feature. We thus pose the problem instead using BPDN [6] (7)

$$\min_{\mathbf{a}\in\mathbb{R}^N} \|\mathbf{a}\|_1 \quad \text{subject to} \quad \|\mathbf{\Phi y} - \mathbf{\Phi D a}\|_2^2 \leq \epsilon^2. \tag{22}$$

or as the LASSO [41] (8)

$$\min_{\mathbf{a}\in\mathbb{R}^N} \|\mathbf{\Phi y} - \mathbf{\Phi D a}\|_2^2 \quad \text{subject to} \quad \|\mathbf{a}\|_1 \leq \rho. \tag{23}$$

Solving these can produce an informative representation using few features instead of an exact fit by many.

Using features with dimensions mapped to $[0, 1]$, and a column-normalized dimension-reduced dictionary $\mathbf{\Phi D}$, Fig. 5(a) shows the results of using BPDN (22); and Fig. 5(b) shows the results when we pose the problem as the LASSO (23). (We show only the results from the Lyon model since the other features did not give significantly different results.) In both cases, we use SGPL1 [2] with

[4] Personal communication.

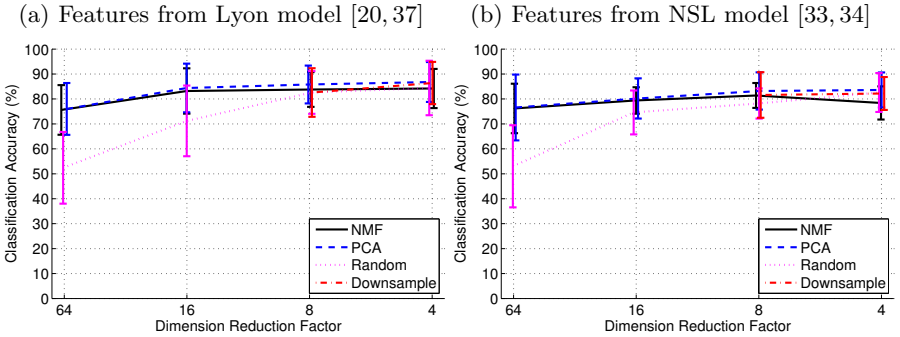


Fig. 6. Mean classification accuracy (5 classes) of SRC based on the LASSO (23) with $\rho := 1$, with features having dimensions mapped to $[0, 1]$ and a normalized projected dictionary, for ATM features derived from a larger frequency range, four dimension reduction methods, and several reduction factors. Overlaid is the standard deviation.

at most 100 iterations, and use the result whether it is in the feasible set or not. This is different from our approach to solving (21), where we run ℓ_1-Magic [5] if SPGL1 fails, and then use the minimum ℓ_2-norm solution if this too fails. In our experiments, we see (23) is solved nearly all the time for $\rho := 1$, and (22) is solved only about 5% of the time with $\epsilon^2 := 10^{-4}$; yet we see no significant differences between the accuracies of both cases. With these changes, we see a slight increase in mean accuracies to about 68% for the features derived from the Lyon model [20, 37], but still far from the 91% reported in [31].

The third significant problem comes from the definition of the features. We find that accuracy improves slightly if we use features from a wider frequency range than the four octaves mentioned in [31], e.g., all 86 bands of the AS from the Lyon model, covering 80 – 7630 Hz [20, 37], or all 128 bands of the AS from the NSL model [33, 34], logarithmically spread over 180 – 7246 Hz. With these changes, however, our mean accuracies do not exceed 70%.

The only way we have found to obtain something close to the 91% mean accuracy reported in [31] is to limit the classification problem to the first five genres of GTZAN: blues, classical, country, disco, and hiphop. Figure 6 shows our results using features derived from the Lyon and NSL models with a wide-frequency range, dimensions mapped to $[0, 1]$, and solving the problem posed with LASSO (23). Though we see the standard deviations are smaller, we still cannot say one feature reduction method performs signficantly different than any other, in contradiction to the findings of [31].

5 Conclusion

Were the difficult problem of music genre recognition solved, it would present a wonderful tool for exploring many interesting questions; and were it solved using solely acoustic features, it would say something significant about a process that appears influenced by much more than sound. Though the approach and results of [31] appear extremely promising in light of state of the art — it is based on a perceptually-informed acoustic feature and a classification method built upon sparse representations in exemplars, which has its own biological

motivations, e.g., [19, 18] — we have not been able to reproduce their results without reducing the number of classes from 10 to 5. We have shown in as much detail possible the variety of decisions we have had to make in our work, and provide all our code for independent verification: `http://imi.aau.dk/~bst/software/`. Though our results point to a negative conclusion with regard to [31], we have confirmed the observation of [47] that the performance of SRC appears robust to the features used. We have found evidence that features modeled on the primary auditory cortex do not perform significantly different from a feature that is not perceptually based. Indeed, it does not make sense to us why perceptually-based features would be more discriminative for the recognition of genre. Finally, we have also shown that relaxing the constraints in the sparse representation component of SRC improves classification accuracy.

As a postscript, we have found in further work [39] that we can increase the mean accuracy of SRC with ATM in music genre recognition to 82% using the Lyon model and downsampling the AS by a factor of 40 (from 22,050 Hz to 551 Hz) before performing the modulation analysis. This performance increase, however, appears irrelevant with respect to genre recognition. When we look beyond the summary statistics, we see this method confidently applies quite illogical classifications, e.g., "Why?" by Bronski Beat is supposedly Classical. We find that its results are highly sensitive to equalization of the audio, and it can be made to label the same piece of music differently if we shape the spectrum in minor ways. Furthermore, we find that the music this method claims is highly representative of a specific genre is not similarly labeled by a listener able to recognize the same genre. Thus, SRC with ATM appears to be choosing labels based on confounding factors of genre. Our future work aims at determining these factors.

Acknowledgments

B. L. Sturm is supported in part by Independent Postdoc Grant 11-105218 from Det Frie Forskningsrd. The authors would like to thank Dr. Costas Kotropoulos and Yannis Panagakis for the helpful discussions about their work.

References

1. Baumann, S., Pohle, T., Vembu, S.: Towards a socio-cultural compatibility of MIR systems. In: Proc. ISMIR. pp. 460–465. Barcelona, Spain (Oct 2004)
2. van den Berg, E., Friedlander, M.P.: Probing the pareto frontier for basis pursuit solutions. SIAM J. on Scientific Computing 31(2), 890–912 (Nov 2008)
3. Bergstra, J., Casagrande, N., Erhan, D., Eck, D., Kégl, B.: Aggregate features and adaboost for music classification. Machine Learning 65(2-3), 473–484 (June 2006)
4. Bingham, E., Mannila, H.: Random projection in dimensionality reduction: Application to image and text data. In: Proc. Int. Conf. Knowledge Discovery Data Mining. pp. 245–250. San Francisco, CA (Aug 2001)
5. Candès, E., Romberg, J.: ℓ_1-magic: Recovery of sparse signals via convex programming. Tech. rep., Caltech, Pasadena, CA, USA (2005)
6. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic decomposition by basis pursuit. SIAM J. Sci. Comput. 20(1), 33–61 (Aug 1998)

7. Dasgupta, S.: Experiments with random projection. In: Proc. Conf. Uncertainty in Artificial Intelligence. pp. 143–151. Stanford, CA, USA (June 2000)
8. Davis, G., Mallat, S., Avellaneda, M.: Adaptive greedy approximations. J. Constr. Approx. 13(1), 57–98 (Jan 1997)
9. Fabbri, F.: A theory of musical genres: Two applications. In: Proc. First International Conference on Popular Music Studies. Amsterdam, The Netherlands (1980)
10. Figueiredo, M., Nowak, R., Wright, S.J.: Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. IEEE J. Sel. Topics Signal Process. 1(4), 586–597 (Dec 2007)
11. Gemmeke, J., ten Bosch, L., L.Boves, Cranen, B.: Using sparse representations for exemplar based continuous digit recognition. In: Proc. EUSIPCO. pp. 1755–1759. Glasgow, Scotland (Aug 2009)
12. Giacobello, D., Christensen, M., Murthi, M.N., Jensen, S.H., Moonen, M.: Enhancing sparsity in linear prediction of speech by iteratively reweighted ℓ_1-norm minimization. In: Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process. Dallas, TX (Mar 2010)
13. Gjerdingen, R.O., Perrott, D.: Scanning the dial: The rapid recognition of music genres. J. New Music Research 37(2), 93–100 (Spring 2008)
14. Grant, M., Boyd, S.: CVX: Matlab software for disciplined convex programming, version 1.21. `http://cvxr.com/cvx` (Apr 2011)
15. Greenberg, S., Kingsbury, B.E.D.: The modulation spectrogram: in pursuit of an invariant representation of speech. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. pp. 1647–1650. Munich, Germany (Apr 1997)
16. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE 86(11), 2278–2324 (Nov 1998)
17. Lena, J.C., Peterson, R.A.: Classification as culture: Types and trajectories of music genres. American Sociological Review 73, 697–718 (Oct 2008)
18. Lewicki, M.S.: Efficient coding of natural sounds. Nature Neuroscience 5(4), 356–363 (Mar 2002)
19. Lewicki, M.S., Sejnowski, T.J.: Learning overcomplete representations. Neural Computation 12, 337–365 (Feb 2000)
20. Lyon, R.F.: A computational model of filtering, detection, and compression in the cochlea. In: Proc. ICASSP. pp. 1282–1285 (1982)
21. Majumdar, A., Ward, R.K.: Robust classifiers for data reduced via random projections. IEEE Trans. Systems, Man, Cybernetics 40(5), 1359–1371 (Oct 2010)
22. Mayer, R., Neumayer, R., Rauber, A.: Rhyme and style features for musical genre classification by song lyrics. In: Proc. Int. Symp. Music Info. Retrieval (2008)
23. McKay, C., Fujinaga, I.: Music genre classification: Is it work pursuing and how can it be improved? In: Proc. Int. Symp. Music Info. Retrieval (2006)
24. Mesgarani, N., Slaney, M., Shamma, S.A.: Discrimination of speech from nonspeech based on multiscame spectro-temporal modulations. IEEE Trans. Audio, Speech, Lang. Process. 14(3), 920–930 (May 2006)
25. Mitra, S.K.: Digital Signal Processing: A Computer Based Approach. McGraw Hill, 3 edn. (2006)
26. Pachet, F., Cazaly, D.: A taxonomy of musical genres. In: Proc. Content-based Multimedia Information Access Conference. Paris, France (Apr 2000)
27. Pampalk, E., Flexer, A., Widmer, G.: Hierarchical organization and description of music collections at the artist level. In: Research and Advanced Technology for Digital Libraries. pp. 37–48 (2005)
28. Panagakis, Y., Benetos, E., Kotropoulos, C.: Music genre classification: A multilinear approach. In: Proc. ISMIR. pp. 583–588. Philadelphia, PA (Sep 2008)

29. Panagakis, Y., Kotropoulos, C.: Music genre classification via topology preserving non-negative tensor factorization and sparse representations. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. pp. 249–252. Dallas, TX (Mar 2010)
30. Panagakis, Y., Kotropoulos, C., Arce, G.R.: Music genre classification using locality preserving non-negative tensor factorization and sparse representations. In: Proc. Int. Symp. Music Info. Retrieval. pp. 249–254. Kobe, Japan (Oct 2009)
31. Panagakis, Y., Kotropoulos, C., Arce, G.R.: Music genre classification via sparse representations of auditory temporal modulations. In: Proc. European Signal Process. Conf. Glasgow, Scotland (Aug 2009)
32. Panagakis, Y., Kotropoulos, C., Arce, G.R.: Non-negative multilinear principal component analysis of auditory temporal modulations for music genre classification. IEEE Trans. Acoustics, Speech, Lang. Process. 18(3), 576–588 (Mar 2010)
33. Ru, P.: Cortical Representations and Speech Recognition. Ph.D. thesis, University of Maryland, College Park, MD, USA (Dec 1999)
34. Ru, P.: Multiscale multirate spectro-temporal auditory model. Tech. rep., Neural Systems Laboratory, University of Maryland College Park (2001), `http://www.isr.umd.edu/Labs/NSL/Software.htm`
35. Sainath, T.N., Carmi, A., Kanevsky, D., Ramabhadran, B.: Bayesian compressive sensing for phonetic classification. In: Proc. ICASSP (2010)
36. Shamma, S.A.: Encoding sound timbre in the auditory system. IETE J. Research 49(2), 145–156 (Mar-Apr 2003)
37. Slaney, M.: Auditory toolbox. Tech. rep., Interval Research Corporation (1998)
38. Sordo, M., Celma, O., Blech, M., Guaus, E.: The quest for musical genres: Do the experts and the wisdom of crowds agree? In: Proc. ISMIR (2008)
39. Sturm, B.L.: Three revealing experiments in music genre recognition. In: Proc. Int. Soc. Music Info. Retrieval. Porto, Portugal (Oct submitted 2012)
40. Sukittanon, S., Atlas, L.E., Pitton, J.W.: Modulation-scale analysis for content identification. IEEE Trans. Signal Process. 52(10), 3023–3035 (Oct 2004)
41. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. Royal Statist. Soc. B 58(1), 267–288 (Jan 1996)
42. Tropp, J.A., Wright, S.J.: Computational methods for sparse solution of linear inverse problems. Proc. IEEE 98(6), 948–958 (June 2010)
43. Tzanetakis, G., Cook, P.: Musical genre classification of audio signals. IEEE Trans. Speech Audio Process. 10(5), 293–302 (July 2002)
44. Wang, K., Shamma, S.A.: Spectral shape analysis in the central auditory system. IEEE Trans. Speech Audio Process. 3(5), 382–395 (Sep 1995)
45. Woolley, S.M.N., Fremouw, T.E., Hsu, A., Theunissen, F.E.: Tuning for spectro-temporal modulations as a mechanishm for auditory discrimination of natural sounds. Nature Neuroscience 8(10), 1371–1379 (Oct 2005)
46. Wright, J., Ma, Y., Mairal, J., Sapiro, G., Huang, T., Yan, S.: Sparse representation for computer vision and pattern recognition. Proc. IEEE 98(6), 1031–1044 (June 2009)
47. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. IEEE Trans. Pattern Anal. Machine Intell. 31(2), 210–227 (Feb 2009)
48. Yang, A.Y., Ganesh, A., Zhou, Z., Sastry, S.S., Ma, Y.: A review of fast l1-minimization algorithms for robust face recognition. (preprint) (2010), `http://arxiv.org/abs/1007.3753`
49. Yang, X., Wang, K., Shamma, S.A.: Auditory representations of acoustic signals. IEEE Trans. Info. Theory 38(2), 824–839 (Mar 1992)

A Survey of Music Recommendation Systems and Future Perspectives

Yading Song, Simon Dixon, and Marcus Pearce *

Centre for Digital Music
Queen Mary University of London
{yading.song, simon.dixon, marcus.pearce}@eecs.qmul.ac.uk

Abstract. Along with the rapid expansion of digital music formats, managing and searching for songs has become significant. Though music information retrieval (MIR) techniques have been made successfully in last ten years, the development of music recommender systems is still at a very early stage. Therefore, this paper surveys a general framework and state-of-art approaches in recommending music. Two popular algorithms: collaborative filtering (CF) and content-based model (CBM), have been found to perform well. Due to the relatively poor experience in finding songs in long tail and the powerful emotional meanings in music, two user-centric approaches: context-based model and emotion-based model, have been paid increasing attention. In this paper, three key components in music recommender - user modelling, item profiling, and match algorithms are discussed. Six recommendation models and four potential issues towards user experience, are explained. However, subjective music recommendation system has not been fully investigated. To this end, we propose a motivation-based model using the empirical studies of human behaviour, sports education, music psychology.

Keywords: Music recommendation; music information retrieval; collaborative filtering; content-based model; emotion-based model; motivation-based model; music psychology

1 Introduction

With the explosion of network in the past decades, internet has become the major source of retrieving multimedia information such as video, books, and music etc. People has considered that music is an important aspect of their lives and they listen to music, an activity they engaged in frequently. Previous research has also indicated that participants listened to music more often than any of the other activities [57] (i.e. watching television, reading books, and watching movies). Music, as a powerful communication and self-expression approach, therefore, has appealed a wealth of research.

* Yading is supported by the China Scholarship Council. We would like to thank Geraint A. Wiggins for his advices.

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

However, the problem now is to organise and manage the million of music titles produced by society [51]. MIR techniques have been developed to solve problems such as genre classification [42, 75], artist identification [46], and instrument recognition [49]. Since 2005, an annual evaluation event called Music Information Retrieval Evaluation eXchange (MIREX[1]) is held to facilitate the development of MIR algorithms.

Additionally, music recommender is to help users filter and discover songs according to their tastes. A good music recommender system should be able to automatically detect preferences and generate playlists accordingly. Meanwhile, the development of recommender systems provides a great opportunity for industry to aggregate the users who are interested in music. More importantly, it raises challenges for us to better understand and model users' preferences in music [76].

Currently, based on users' listening behaviour and historical ratings, collaborative filtering algorithm has been found to perform well [9]. Combined with the use of content-based model, the user can get a list of similar songs by low-level acoustic features such as rhythm, pitch or high-level features like genre, instrument etc [7].

Some music discovery websites such as Last.fm[2], Allmusic[3], Pandora[4] and Shazam[5] have successfully used these two approaches into reality. At the meantime, these websites provide an unique platform to retrieve rich and useful information for user studies.

Music is subjective and universal. It not only can convey emotion, but also can it modulate a listener's mood [23]. The tastes in music are varied from person to person, therefore, the previous approaches cannot always meet the users' needs. An emotion-based model and a context-based model have been proposed [18, 34]. The former one recommends music based on mood which allows the user to locate their expected perceived emotion on a 2D valence-arousal interface [22]. The latter one collects other contextual information such as comments, music review, or social tags to generate the playlist. Though hybrid music recommender systems would outperform the conventional models, the development is still at very early stage [88]. Due to recent studies in psychology, signal processing, machine learning and musicology, there is much room for future extension.

This paper, therefore, surveys a general music recommender framework from user profiling, item modelling, and item-user profile matching to a series of state-of-art approaches. Section 2 gives a brief introduction of components in music recommendation systems and in section 3, the state-of-art recommendation techniques are explained. To the end of this paper, we conclude and propose a new model based on users' motivation.

[1] http://www.music-ir.org/mirex/wiki/MIREX_HOME
[2] http://www.last.fm/
[3] http://www.allmusic.com/
[4] http://www.pandora.com
[5] http://www.shazam.com/

2 Components in Music Recommender System

Generally, a music recommender system consists of three key components - users, items and user-item matching algorithms. User profiling (see section 2.1) addresses the variation in users' profile. This step aims at differentiating their music tastes using basic information. Item profiling (see section 2.2) on the contrary, describes three different types of metadata - editorial, cultural and acoustic, which are used in different recommendation approaches. In section 2.3, we explain the query in music recommender systems, and the matching algorithms are presented in section 3.

2.1 User Modelling

A successful music recommender needs to meet users' various requirements. However, obtaining user information is expensive in terms of financial costs and human labor [74]. For user-oriented design, lots of efforts on user studies need to be investigated.

User modelling, as the one of the key elements, it models the difference in profile. For example, the difference in geographic region or age, their music preferences might be different. Interestingly, other factors such as gender, life styles, and interests could also determine their choices of music.

Recent research has revealed that intelligence, personality and the users' preference in music are linked [57]. According to Rentfrow and Gosling [26, 58] who had investigated the relationship between music preference and Big-Five Inventory (BFI: openness, conscientiousness, extraversion, agreeableness, and neuroticism), their studies showed a highly extraverted person would tend to choose the music which is energetic, while a greater preference for rhythmic and energetic music was associated with greater extraversion and agreeableness. User modelling, therefore, is essential in prediction of their music taste. It has been divided into two parts: user profile modelling and user experience modelling.

First Step - User Profile Modelling Celma [14] suggested that the user profile can be categorised into three domains: *demographic*, *geographic*, and *psychographic* (shown in Table 1). Based on the steadiness, psychological data has been further divided into stable attributes which are essential in making a long term prediction and fluid attributes which can change on an hour to hour basis [24].

Data type	Example
Demographic	Age, marital status, gender etc.
Geographic	Location, city, country etc.
Psychographic	*Stable*: interests, lifestyle, personality etc. *Fluid*: mood, attitude, opinions etc.

Table 1. User profile classification

Second Step - User Listening Experience Modelling Depending on the level of music expertise, their expectations in music are varied accordingly. Jennings [32] analysed the different types of listeners whose age range from 16-45 and categorised the listeners into four groups: *savent*, *enthusiasts*, *casuals*, *indifferents* (see Table 2).

Type	Percentage	Features
Savants	7	Everything in life seems to be tied up with music. Their musical knowledge is very extensive.
Enthusiasts	21	Music is a key part of life but is also balanced by other interests.
Casuals	32	Music plays a welcome role, but other things are far more important.
Indifferents	40	They would not lose much sleep if music ceased to exist, they are a predominant type of listeners of the whole population.

Table 2. Use listening experience categorisation

This information gives us a good example that their expertise needs to be considered when designing user-oriented recommendation systems. For instance, based on their expectation, we need to consider the amount of music to be discovered and filtered in long tail which represents interesting and unknown music but hidden in the tail of the popularity curve [3]. Other user information including access pattern, listening behaviour are also useful for user modelling and dynamic optimisation [50]. Exploring user information can be either done through the initial survey or observing their behaviour of music in long tail.

2.2 Item Profiling

The second component of recommender systems is music item. It defines a various of information that used in MIR. In 2005, Pachet [53] classified the music metadata into three categories: *editorial metadata* (EM), *cultural metadata* (CM), and *acoustic metadata* (AM).

- **Editorial metadata**: Metadata obtained by a single expert or group of experts. This is obtained literally by the editor, and also it can be seen as the information provided by them. E.g. the cover name, composer, title, or genre etc.
- **Cultural metadata**: Metadata obtained from the analysis of corpora of textual information, usually from the Internet or other public sources. This information results from an analysis of emerging patterns, categories or associations from a source of documents. E.g. Similarity between music items.
- **Acoustic metadata**: Metadata obtained from an analysis of the audio signal. This should be without any reference to a textual or prescribed information. E.g. Beat, tempo, pitch, instrument, mood etc.

Editorial metadata are mostly used in metadata information retrieval (see section 3.1), and cultural metadata have been largely used in context-based information retrieval (see section 3.5). However, most music recommendation systems are using acoustic metadata for discovering music which is named as content-based information retrieval (see section 3.3).

2.3 Query Type

Assuming that the users have already known the information about the music, the quickest way to search for music is via key editorial information such as title, the name of the singer and lyrics etc. However, it is not always the case of knowing them. In the past ten years, an advanced and more flexible music information retrieval system called "query by humming/singing system (QBSH)" was developed [25]. It allows the user to find the songs either by humming or singing.

Nevertheless, it still requires lots of human efforts. In recommender systems, a more appropriate way is to use listening histories or seed music as the query to detect their music preferences.

3 State-of-art Approaches in Music Recommendation

An ideal music recommender system should be able to automatically recommend personalised music to human listeners [36, 52]. Different from books or movies, the length of a piece of music is much shorter, and the times that listening their favourite songs are normally more than once.

The existing recommender systems such as *Amazon*, *Ebay* have gained a great success. It can recommend complementary goods, the buyer can compare the products (new-item/old-item) and negotiate with the sellers [69]. However, music recommender is not only giving products with reasonable price, but suggesting them personalised music.

So far, many music discovery websites such as *Last.fm*, *Allmusic*, *Pandora*, *Audiobaba*[6], *Mog*[7], *Musicovery* [8], *Spotify*[9], *Apple "Genius"* have aggregated millions of users, and the development is explosive [10, 11]. In this section, we present the most popular approaches, metadata information retrieval (see section 3.1), collaborative filtering (see section 3.2), content-based information retrieval (see section 3.3), emotion-based model (see section 3.4), context-based information retrieval (see section 3.5) and hybrid models (see section 3.6). At the end of each approach, their limitations are described.

3.1 Metadata Information Retrieval (Demographic-based Model)

As the most fundamental method, it is the easiest way to search for music. Metadata information retrieval uses textual metadata (editorial information)

[6] http://audiobaba.com/
[7] http://www.mog.com/
[8] http://www.musicovery.com/
[9] http://www.spotify.com/

supplied by the creators, such as the title of the song, artist name, and lyrics to find the target songs [20].

Limitation Though it is fast and accurate, the drawbacks are obvious. First of all, the user has to know about the editorial information for a particular music item. Secondly, it is also time consuming to maintain the increasing metadata. Moreover, the recommendation results is relatively poor, since it can only recommend music based on editorial metadata and none of the users' information has been considered.

3.2 Collaborative Filtering

To recommend items via the choice of other similar users, collaborative filtering technique has been proposed [28]. As one of the most successful approaches in recommendation systems, it assumes that if user X and Y rate n items similarly or have similar behaviour, they will rate or act on other items similarly [59].

Instead of calculating the similarity between items, a set of 'nearest neighbour' users for each user whose past ratings have the strongest correlation are found. Therefore, scores for the unseen items are predicted based on a combination of the scores known from the nearest neighbours [65]. Collaborative filtering is further divided into three subcategories: *memory-based*, *model-based*, and *hybrid* collaborative filtering [63, 68].

Memory-based Collaborative Filtering Memory-based collaborative filtering is to predict the item based on the entire collections of previous ratings. Every user is grouped with people with similar interests, so that a new item is produced by finding the nearest neighbour using a massive number of explicit user votes [9].

Model-based Collaborative Filtering In contrast to memory-based CF, model-based CF uses machine learning and data mining algorithms which allow the system to train and model the users' preferences. It represents the user preference by a set of rating scores and constructs a special prediction model [2]. Based on the known model, the system makes prediction for test and real-world data.

Hybrid Collaborative Filtering Hybrid CF model is to make prediction by combining different CF models. It has been proved that hybrid CF model outperforms any individual method [83].

Limitations Because of the subjectivity in music, the assumption that users with similar behaviours may have same tastes has not been widely studied. Though collaborative filtering recommender works well, the key problems such as cold start, popularity bias are unavoidable [27].

- **Popularity bias** Generally, popular music can get more ratings. The music in long tail, however, can rarely get any. As a result, collaborative filtering mainly recommend the popular music to the listeners. Though giving popular items are reliable, it is still risky, since the user rarely get pleasantly surprised.
- **Cold start** It is also known as data sparsity problems. At an early stage, few ratings is provided. Due to the lack of these ratings, prediction results are poor.
- **Human effort** A perfect recommender system should not involve too much human efforts, since the users are not always willing to rate. The ratings may also grow towards those who do rate, but it may not be representative. Because of this absence of even distributed ratings, it can either give us false negative or false positive results.

3.3 Content/Audio/Signal-based Music Information Retrieval

Different from collaborative filtering technique, content-based approach makes prediction by analysing the song tracks [2, 41]. It is rooted in information retrieval and information filtering [13] that recommends a song which is similar to those the user has listened to in the past rather than what the user have rated 'like' [4, 43]. Lots of research have been paid attention on extracting and comparing the acoustic features in finding perceptual similar tracks [8, 45]. The most representative ones so far are timbre, rhythm [7, 10, 11].

Based on the extracted features, the distance between songs are measured [43]. Three typical similarity measurements are listed below [44].

- **K-means clustering with Earth-Mover's Distance**: It computes a general distance between Gaussian Mixture Models (GMM) by combining individual distance between gaussian components [62].
- **Expectation-Maximization with Monte Carlo Sampling**: This measurement makes use of vectors sampled directly from the GMMs of the two songs to be compared; the sampling is performed computationally via random number generation [51].
- **Average Feature Vectors with Euclidean Distance**: It calculates low-order statistics such as mean and variance over segments [16].

Query by Humming (QBSH) Humming and singing are the natural way to express the songs [31]. In the early 1990s, based on content-based model, query by humming system was proposed [25, 80]. Early query by humming systems were using melodic contour which had been seen as the most discriminative features in songs.

It follows three steps: construction of the songs database, transcription of the users' melodic information query and pattern matching algorithms which are used to get the closest results from collections [1]. In the past few years, except melody, a better performance has also been achieved by embedding with lyrics and enhancing the main voice [19, 77].

Limitations To some extent, content-based model solves the problems in collaborative filtering. For instance, by measuring the similarity of acoustic features between songs, the system can recommend music using distance measurements. Therefore, no human rating is needed. However, similarity-based method has not been fully investigated in terms of listeners' preference. None of the research proved that similar behaviour leads to the choice of same music.

Since content-based model largely depends on acoustic features, the number of selected features needs to be further considered. Moreover, other user information and non-acoustic information should be included for future modification and augmentation.

3.4 Emotion-based Model

Music as a self-expression tool, it always performs with affection. Rich in content and expressivity [86], the conventional approaches for music information retrieval are no longer sufficient. Music emotion has appealed lots of research and it has become the main trend for music discovery and recommendation [34]. A commercial web service called '*Musicovery*' uses the fundamental emotion model (2D valence-arousal) found by psychologists. It allows users to locate their expected perceived emotion in a 2D space: *valence* (how positive or negative) and *arousal* (how exciting or calming).

Similar to content-based model, the emotion perception is associated with different patterns of acoustic cues [6, 35, 48, 61]. Different perceptual features such as energy, rhythm, temporal, spectral, and harmony have been widely used in emotion recognition [84].

Limitations

- **Data collection** In order to accurately model the system, a great amount of dataset are needed. However, finding the reliable ground truth is expensive and requires a lot of human efforts [67]. Instead of human annotation [73], social tags [36, 74], annotation games like *MajorMiner* [47] and *TagATune* [37], lyrics or music review are being used for data collection.
- **Ambiguity and granularity** Emotion itself is hard to define and describe. The same affective feeling experienced by different people may give different emotion expression (i.e. cheerful, happy) and there is no perfect relationship between affective terms with emotions [64, 85]. Some research were based on basic taxonomy (sad, happy, angry etc.), but it cannot describe the richness of our human perception. MIREX evaluation has categorised emotion into 5 mood clusters [29] (see Table 3). Russell [60] found a circumflex model which affective concepts fall in a circle in the following order: pleasure (0°), excitement (45°), arousal (90°), distress (135°), displeasure (180°), depression (225°), sleepiness (270°), and relaxation (315°). It can represent the structure of affective experience and now it has been become the most noted 2D valence-arousal emotion model. The problem of classifying emotion, therefore has been solved, since each point on the plane represents an affective term.

Cluster 1	Passionate, rousing, confident, boisterous, rowdy
Cluster 2	Rollicking, cheerful, fun, sweet, amiable/good natured
Cluster 3	Literate, poignant, wistful, bittersweet, autumnal, brooding
Cluster 4	Humorous, silly, campy, quirky, whimsical, witty, wry
Cluster 5	Aggressive, fiery, tense/anxious, intense, volatile, visceral

Table 3. MIREX five mood categories

3.5 Context-based Information Retrieval

Rather than using acoustic features in content-based model and ratings in collaborative filtering, context-based information retrieval model uses the public opinion to discover and recommend music [18]. Along with the development of social networks such as *Facebook*[10], *Youtube*[11], and *Twitter*[12], these websites provide us rich human knowledge such as comments, music review, tags and friendship networks [36].

Context-based information retrieval, therefore, uses web/document mining techniques to filter out important information to support problems like artist similarity, genre classification, emotion detection [82], semantic space [39, 40] etc. Some researchers have suggested that the use of social information has outperformed content-based model [70, 81].

However, the same problems as collaborative filtering, the popular music can always get more public opinions than those in long tail [21]. Eventually, rich music gets richer feedback, it again results in a popularity bias problem.

3.6 Hybrid Model Information Retrieval

Hybrid model aims at combining two or more models to increase the overall performance. Burke [9] pointed out several methods to build a hybrid model such as *weighted*, *switching*, *mixed*, *feature combination*, and *cascade*. There is no doubt that a proper hybrid model would outperform a single approach, since it can incorporate the advantages of both methods while inheriting the disadvantages of neither [65, 87, 88].

3.7 Other Issues

We have discussed above the essential problems in music recommender systems, the other issues such as dynamic evolvement, playlist generation, user interface design and evaluation need to be further considered. Though it doesn't affect the recommendation performance, it certainly influence the user listening experience.

[10] https://www.facebook.com/
[11] http://www.youtube.com/
[12] https://twitter.com/

Dynamic Evolvement As the users aggregate in the recommender systems, it needs to be able to adapt to new data such as user listening histories and listening behaviour to further personalised their music taste [30]. This procedure is called evolvement. It addresses the problem that when the new user comes and new items into the system, it can dynamically and automatically evolve itself [65].

Playlist Generation Another issue is the sequence of the playlist [38]. Most of the recommender systems are not flexible, because the playlist is ordered by the similarity distance between seed songs. Though the most similar songs are given in order, the theme and mood can be dramatically changed in between. This may result in the dissatisfaction and discontinuation of the songs.

Research indicates that a playlist should have a main theme (mood, event, activity) evolve with time [17]. Rather than randomly shuffling, human skipping behaviour can be considered for dynamic playlist generation [15, 54]. For example, assuming that the users dislike the song when they skipped it, the system therefore, removes the songs which are similar to the song which they skipped [55, 56, 78].

User Interface Design A bad design of user interface cannot affect the accuracy of the system, it does influence the ratings and listening experience. A clear design always gives the user a better understanding of the system. Moreover, an overall control of the system and less human efforts required for operation should be considered during designing.

Evaluation There is no common objective evaluation in music recommendation systems [72]. Most of the evaluation techniques are based on subjective system testing which let users to rank the systems given the playlist generated by different approaches [5, 79]. However, it is very expensive in terms of financial costs and human labor. Another important factor is that the evaluation in different regions (i.e. different background, age, language) might give different results. Hence, a proper evaluation criteria is essential and highly recommended.

4 Conclusion and Future Work

In this paper, we explain a basic metadata-based model and two popular music recommender approaches: collaborative filtering and content-based model. Though they have achieved great success, their drawbacks such as popularity bias and human efforts are obvious. Moreover, the use of hybrid model would outperform a single model since it incorporates the advantages of both methods. Its complexity is not fully studied yet.

Due to the subjective nature in music and the issues existing in the previous methods, two human-centred approaches are proposed. By considering affective and social information, emotion-based model and context-based model largely improved the quality of recommendation. However, this research is still at an early stage.

As we can see from the development of music recommenders over the past years, the given results tend to be more personalised and subjective. Only considering the music itself and human ratings are no longer sufficient. A great amount of work in recent years have been done in music perception, psychology, neuroscience and sport which study the relationship between music and the impact of human behaviour. David Huron also mentioned music has sex and drug-like qualities. Undoubtably, music always has been an important component of our life, and now we have greater access to it.

Researches in psychology pointed out that music not only improves mood, increases activation, visual and auditory imagery, but also recalls of associated films or music videos and relieves stress [33]. Moreover, the empirical experiments in sport mentioned that the main benefits for listening to the music which include work output extension, performance enhancement, and dissociation from unpleasant feelings etc [71]. For example, athletes prefer uptempo, conventional, intense, rebellious, energetic, and rhythmic music rather than reflective and complex music [66]. An important fact found by psychologists is that users' preference in music is linked to their personality. Also worth mentioning that fast, upbeat music produces a stimulative effect whereas slow, while soft music produces a sedative effects [12]. All of these highlight that music recommender is not only a tool for relaxing, but also acts as an effective tool to meet our needs under different contexts. To our knowledge, there is few research based on these empirical results.

Designing a personalised music recommender is complicated, and it is challenging to thoroughly understand the users' needs and meet their requirements. As discussed above, the future research direction will be mainly focused on user-centric music recommender systems. A survey among athletes showed practitioners in sport and exercise environments tend to select music in a rather arbitrary manner without full consideration of its motivational characteristics. Therefore, future music recommender should be able to lead the users reasonably choose music. To the end, we are hoping that through this study we can build the bridge among isolated research in all the other disciplines.

Acknowledgments. Yading is supported by the China Scholarship Council. We would like to thank Geraint A. Wiggins, Steven Hargreaves and Emmanouil Benetos for their advices.

References

1. Ricardo A. Baeza-Yates and Chris H. Perleberg. Fast and Practical Approximate String Matching. In *Combinatorial Pattern Matching, Third Annual Symposium*, pages 185–192, 1992.
2. G. Adomavicius and A. Tuzhilin. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-art and Possible Extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, June 2005.
3. C Anderson. *The Long Tail. Why the Future of Business is selling less of more.* Hyperion Verlag, 2006.

4. Jean-julien Aucouturier and Francois Pachet. Music Similarity Measures: What is the Use. In *Proceedings of the ISMIR*, pages 157–163, 2002.
5. Luke Barrington, Reid Oda, and G. Lanckriet. Smarter Than Genius? Human Evaluation of Music Recommender Systems. In *10th International Society for Music Information Retrieval Conference*, number ISMIR, pages 357–362, 2009.
6. Kerstin Bischoff, Claudiu S Firan, Raluca Paiu, Wolfgang Nejdl, L S De, Cyril Laurier, and Mohamed Sordo. Music Mood and Theme Classification - A Hybrid Approach. In *10th International Society for Music Information Retrieval Conference*, number Ismir, pages 657–662, 2009.
7. Dmitry Bogdanov and Perfecto Herrera. How Much Metadata Do We Need in Music Recommnendation? A Subjective Evaluation Using Preference Sets. In *12th International Society for Music Information Retrieval Conference*, number ISMIR 2011, pages 97–102, 2011.
8. Dmitry Bogdanov, J. Serra, Nicolas Wack, Perfecto Herrera, and Xavier Serra. Unifying Low-level and High-level Music Similarity Measures. *IEEE Transactions on Multimedia*, 13(99):1–1, 2011.
9. Robin Burke. Hybrid Recommender Systems: Survey and Experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, 2002.
10. Pedro Cano, Markus Koppenberger, and Nicolas Wack. An Industrial-strength Content-based Music Recommendation System. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '05*, page 673, New York, New York, USA, 2005. ACM Press.
11. Pedro Cano, Markus Koppenberger, and Nicolas Wack. Content-based Music Audio Recommendation. In *Proceedings of the 13th annual ACM international conference on Multimedia*, number ACM, pages 211–212, 2005.
12. FRD Carpentier. Effects of Music on Physological Arousal, Exploration into Tempo and Genre. *Media Psychology*, 10(3):339–363, 2007.
13. M.A. Casey, Remco Veltkamp, Masataka Goto, Marc Leman, Christophe Rhodes, and Malcolm Slaney. Content-based Music Information Retrieval: Current Directions and Future Challenges. *Proceedings of the IEEE*, 96(4):668–696, 2008.
14. O. Celma Herrada. Music Recommendation and Discovery in the Long Tail. *PhD Thesis*, 2009.
15. Zeina Chedrawy and S. Abidi. A Web Recommender System for Recommending, Predicting and Personalizing Music Playlists. In *Web Information Systems Engineering-WISE 2009*, pages 335–342. Springer, 2009.
16. Parag Chordia, Mark Godfrey, and Alex Rae. Extending Content-Based Recommendation: The Case of Indian Classical Music. In *ISMIR 2008: proceedings of the 9th International Conference of Music Information Retrieval*, pages 571–576, 2008.
17. Sally Jo Cunningham, David Bainbridge, and Annette Falconer. More of an Art than a Science : Supporting the Creation of Playlists and Mixes. In *Seventh International Conference on Music Information Retrieval*, Victoria, Canada, 2006.
18. M. Datcu, H. Daschiel, A. Pelizzari, M. Quartulli, A. Galoppo, A. Colapicchioni, M. Pastori, K. Seidel, P.G. Marchetti, and S. D'Elia. Music Recommendation Using Content and Context Information Mining. *Geoscience and Remote Sensing, IEEE Transactions on*, 41(12):2923–2936, 2003.
19. C. de la Bandera, A.M. Barbancho, L.J. Tardón, Simone Sammartino, and Isabel Barbancho. Humming Method for Content-Based Music Information Retrieval. *ISMIR 2011*, (Ismir):49–54, 2011.
20. J. Stephen Downie. Music Information Retrieval. *Annual Review of Information Science and Technology*, 37(1):295–340, January 2005.

21. Douglas Eck, Paul Lamere, T. Bertin-Mahieux, and Stephen Green. Automatic Generation of Social Tags for Music Recommendation. *Advances in neural information processing systems*, 20:385–392, 2007.
22. T. Eerola and J. K. Vuoskoski. A Comparison of the Discrete and Dimensional Models of Emotion in Music. *Psychology of Music*, 39(1):18–49, August 2010.
23. Yazhong Feng and Y Zhuang. Popular Music Retrieval by Detecting Mood. In *International Society for Music Information Retrieval 2003*, volume 2, pages 375–376, 2003.
24. Alan Page Fiske, Shinobu Kitayama, Hazel Rose Markus, and R E Nisbett. *The Cultural Matrix of Social Psychology*. 1998.
25. Asif Ghias, Jonathan Logan, David Chamberlin, and Brian C. Smith. Query by Humming. *Proceedings of the third ACM international conference on Multimedia - MULTIMEDIA '95*, pages 231–236, 1995.
26. S Gosling. A Very Brief Measure of the Big-Five Personality Domains. *Journal of Research in Personality*, 37(6):504–528, December 2003.
27. Jonathan L. Herlocker, Joseph a. Konstan, Loren G. Terveen, and John T. Riedl. Evaluating Collaborative Filtering Recommender Systems. *ACM Transactions on Information Systems*, 22(1):5–53, January 2004.
28. Will Hill, Larry Stead, Mark Rosenstein, George Furnas, and South Street. Recommending and Evaluating Choices in a Vitual Community of Use. *Mosaic A Journal For The Interdisciplinary Study Of Literature*, pages 5–12, 1995.
29. X Hu and J. Stephen Downie. Exploring Mood Metadata: Relationships with Genre, Artist and Usage Metadata. In *8th International Conference on Music Information Retrieval*, 2007.
30. Yajie Hu and Mitsunori Ogihara. Nextone Player: A Music Recommendation System Based on User Behavior. In *12th International Society for Music Information Retrieval Conferenceh*, number Ismir, pages 103–108, 2011.
31. Jyh-Shing Roger Jang and Hong-Ru Lee. A General Framework of Progressive Filtering and Its Application to Query by Singing/Humming. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):350–358, February 2008.
32. D Jennings. *Net, Blogs and Rock 'n' Rolls: How Digital Discovery Works and What It Means for Consumers*. 2007.
33. Patrik N Juslin and Daniel Västfjäll. Emotional Responses to Music: the Need to Consider Underlying Mechanisms. *The Behavioral and brain sciences*, 31(5):559–621, October 2008.
34. Y.E. Kim, E.M. Schmidt, Raymond Migneco, B.G. Morton, Patrick Richardson, Jeffrey Scott, J.A. Speck, and Douglas Turnbull. Music Emotion Recognition: A State of the Art Review. In *Proc. of the 11th Intl. Society for Music Information Retrieval (ISMIR) Conf*, number Ismir, pages 255–266, 2010.
35. Fang-Fei Kuo, Meng-Fen Chiang, Man-Kwan Shan, and Suh-Yin Lee. Emotion-based Music Recommendation by Association Discovery from Film Music. In *Proceedings of the 13th annual ACM international conference on Multimedia - MULTIMEDIA '05*, page 507, New York, New York, USA, 2005. ACM Press.
36. Paul Lamere. Social Tagging and Music Information Retrieval. *Journal of New Music Research*, 37(2):101–114, June 2008.
37. Edith L. M. Law, Luis Von Ahn, Roger B. Dannenberg, and Mike Crawford. Tagatune: A Game for Music and Sound Annotation. In *8th International Conference on Music Information Retrieval*, 2007.
38. Jin Ha Lee, Bobby Bare, and Gary Meek. How Similar is too Similar? Exploring Users' Perception of Similarity in Playlist Evaluation. In *International Conference on Music Information Retrieval 2011*, number ISMIR, pages 109–114, 2011.

39. Mark Levy. A Semantic Space for Music Derived from Social Tags. *Austrian Compuer Society*, 1:12, 2007.
40. Mark Levy and Mark Sandler. Music Information Retrieval Using Social Tags and Audio. *IEEE Transactions on Multimedia*, 11(3):383–395, 2009.
41. Qing Li, Byeong Man Kim, Dong Hai Guan, and Duk Oh. A Music Recommender Based on Audio Features. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 532–533, Sheffield, United Kingdom, 2004. ACM.
42. Bass Lines, Emiru Tsunoo, George Tzanetakis, and Nobutaka Ono. Beyond Timbral Statistics : Improving Music Classification Using Percussive. *IEEE Transactions on Audio, Speech and Language Processing*, 19(4):1003–1014, 2011.
43. Beth Logan. Music Recommendation from Song Sets. In *International Conference on Music Information Retrieval 2004*, number October, pages 10–14, Barcelona, Spain, 2004.
44. Terence Magno and Carl Sable. A Comparison of Signal of Signal-Based Music Recommendation to Genre Labels, Collaborative Filtering, Musicological Analysis, Human Recommendation and Random Baseline. In *ISMIR 2008: proceedings of the 9th International Conference of Music Information Retrieval*, pages 161–166, 2008.
45. Chun-man Mak, Tan Lee, Suman Senapati, Yu-ting Yeung, and Wang-kong Lam. Similarity Measures for Chinese Pop Music Based on Low-level Audio Signal Attributes. In *11th International Society for Music Information Retrieval Conference*, number ISMIR, pages 513–518, 2010.
46. M Mandel. Song-level Features and Support Vector Machines for Music Classification. In *Proc. International Conference on Music*, 2005.
47. MI Mandel. A Web-based Game for Collecting Music Metadata. In *In 8th International Conference on Music Information Retrieval (ISMIR)*, 2008.
48. M Mann, TJ Cox, and FF Li. Music Mood Classification of Television Theme Tunes. In *12th International Society for Music Information Retrieval Conference*, number Ismir, pages 735–740, 2011.
49. Janet Marques and Pedro J Moreno. A Study of Musical Instrument Classification Using Gaussian Mixture Models and Support Vector Machines, 1999.
50. M. Ogihara, Bo Shao, Dingding Wang, and Tao Li. Music Recommendation Based on Acoustic Features and User Access Patterns. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(8):1602–1611, November 2009.
51. F. Pachet and J.J. Aucouturier. Improving Timbre Similarity: How High is the Sky? *Journal of negative results in speech and audio sciences*, 1(1):1–13, 2004.
52. François Pachet and Daniel Cazaly. A Taxonomy of Musical Genres. In *Content-Based Multimedia Information Retrieval Access Conference (RIAO)*, number April, 2000.
53. Francois Pachet. Knowledge Management and Musical Metadata. In *Encyclopedia of Knowledge Management*. 2005.
54. Elias Pampalk, Tim Pohle, and Gerhard Widmer. Dynamic Playlist Generation Based on Skipping Behavior. In *Proc. of the 6th ISMIR Conference*, volume 2, pages 634–637, 2005.
55. Steffen Pauws, Berry Eggen, and Miles Davis. PATS : Realization and User Evaluation of an Automatic Playlist Generator PATS : Realization and User Evaluation of an Automatic Playlist Generator. In *3rd International Conference on Music Information Retrieval*, 2002.
56. CBE Plaza. Uncovering Affinity of Artist to Multiple Genres from Social Behavior Data. In *ISMIR 2008: proceedings of the 9th*, pages 275–280, 2008.

57. Peter J. Rentfrow and Samuel D. Gosling. The Do Re Mi's of Everyday Life: The structure and personality correlates of music preferences. *Journal of Personality and Social Psychology*, 84(6):1236–1256, 2003.
58. Peter J Rentfrow and Samuel D Gosling. Message in a Ballad: the Role of Music Preferences in Interpersonal Perception. *Psychological science*, 17(3):236–42, March 2006.
59. Paul Resnick, Hal R Varian, and Guest Editors. Recommender Systems. *Communications of the ACM*, 40(3):56–58, 1997.
60. J.A. Russell. A Circumplex Model of Affect. *Journal of personality and social psychology*, 39(6):1161–1178, 1980.
61. Pasi Saari, Tuomas Eerola, and Olivier Lartillot. Generalizability and Simplicity as Criteria in Feature Selection: Application to Mood Classification in Music. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(99):1–1, 2011.
62. A Salomon. A Content-based Music Similarity Function. In *Cambridge Research Labs-Tech Report*, number June, 2001.
63. Badrul Sarwar, George Karypis, and Joseph Konstan. Item-based Collaborative Filtering Recommendation Algorithms. *Proceedings of the 10th*, pages 285–295, 2001.
64. Bo Shao, Tao Li, and M. Ogihara. Quantify Music Artist Similarity Based on Style and Mood. In *Proceeding of the 10th ACM workshop on Web Information and Data Management*, pages 119–124. ACM, 2008.
65. Yoav Shoham and Marko Balabannovic. Content-Based, Collaborative Recommendation. *Communications of the ACM*, 40(3):66–72, 1997.
66. Stuart D Simpson and Costas I Karageorghis. The Effects of Synchronous Music on 400-m Sprint Performance. *Journal of sports sciences*, 24(10):1095–102, October 2006.
67. Janto Skowronek and M McKinney. Ground Truth for Automatic Music Mood Classification. In *Proc. ISMIR*, pages 4–5, 2006.
68. Xiaoyuan Su and Taghi M. Khoshgoftaar. A Survey of Collaborative Filtering Techniques. *Advances in Artificial Intelligence*, 2009(Section 3):1–19, 2009.
69. Neel Sundaresan. Recommender Systems at the Long Tail. In *of the fifth ACM conference on Recommender systems*, number RecSys 2011, pages 1–5, 2011.
70. Panagiotis Symeonidis, Maria Ruxanda, Alexandros Nanopoulos, and Yannis Manolopoulos. Ternary Semantic Analysis of Social Tags for Personalized Music Recommendation. In *Proc. 9th ISMIR Conf*, pages 219–224. Citeseer, 2008.
71. P.C. Terry and C.I. Karageorghis. Psychophysical Effects of Music in Sport and Exercise: An Update on Theory, Research and Application. In *Proceedings of the 2006 Joint Conference of the Australian Psychological Society and the New Zealand Psychological Society: Psychology Bridging the Tasman: Science, Culture and Practice*, pages 415–419. Australian Psychological Society, 2006.
72. Nava Tintarev and Judith Masthoff. Effective Explanations of Eecommendations: User-centered Design. In *Proceedings of the 2007 ACM conference on Recommender systems*, pages 153–156. ACM, 2007.
73. KTG Tsoumakas and George Kalliris. Multi-Label Classification of Music into Emotions. In *ISMIR 2008: proceedings of the 9th International Conference of Music Information Retrieval*, pages 325–330, 2008.
74. Douglas Turnbull, Luke Barrington, and Gert Lanckriet. Five Approaches to Collecting Tags for Music. In *ISMIR 2008: proceedings of the 9th International Conference of Music Information Retrieval*, pages 225–230, 2008.

75. George Tzanetakis, Student Member, and Perry Cook. Musical Genre Classification of Audio Signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.
76. Alexandra Uitdenbogerd and van Schyndel Ron. A Review of Factors Affecting Music Recommender. In *3rd International Conference on Music Information Retrieval (2002)*, 2002.
77. Erdem Unal, Elaine Chew, Panayiotis G. Georgiou, and Shrikanth S. Narayanan. Challenging Uncertainty in Query by Humming Systems: A Fingerprinting Approach. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):359–371, February 2008.
78. Rob van Gulik and Fabio Vignoli. Visual Playlist Generation on the Artist Map. In *5th International Conference on Music Infomation Retrieval*, number ISMIR2005, pages 520–523, 2005.
79. Fabio Vignoli. A Music Retrieval System Based on User-driven Similarity and its Evaluation. In *International Conference on Music Information Retrieval 2005*, 2005.
80. C.C. Wang, J.S.R. Jang, and Wennen Wang. An Improved Query by Singing/Humming System Using Melody and Lyrics Information. In *11th International Society for Music Information Retrieval Conference*, number Ismir, pages 45–50, 2010.
81. Dingding Wang, Tao Li, and Mitsunori Ogihara. Tags Better Than Audio Features? The Effect of Joint use of Tags and Audio Content Features for Artistic Style Clutering. In *International Conference on Music Information Retrieval 2010*, number ISMIR, pages 57–62, 2010.
82. Ju-chiang Wang, Hung-shin Lee, Hsin-min Wang, and Shyh-kang Jeng. Learning the Similarity of Audio Msuic in Bag-of-Frames Representation from Tagged Music Data. In *International Conference on Music Information Retrieval 2011*, number ISMIR, pages 85–90, 2011.
83. Jun Wang, A.P. De Vries, and M.J.T. Reinders. Unifying User-based and Item-based Collaborative Filtering Approaches by Similarity Fusion Categories. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 501–508. ACM, 2006.
84. Xing Wang, Xiaoou Chen, Deshun Yang, and Yuqian Wu. Music Emotion Classification of Chinese Songs Based on Lyrics using TF*IDF and Rhyme. In *12th International Society for Music Information Retrieval Conference*, number Ismir, pages 765–770, 2011.
85. Dan Yang and W.S. Lee. Disambiguating Music Emotion Using Software Agents. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR04)*, pages 52–58, 2004.
86. Yi-Hsuan Yang. *Music Emotion Recognition*. Tayler and Francis Group, 2011.
87. Kazuyoshi Yoshii, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G Okuno. Hybrid Collaborative and Content-based Music Recommendation Using Probabilistic Model with Latent User Preferences. In *Proceedings of the 7th International Conference on Music Information Retrieval*, pages 296–301, 2006.
88. Kazuyoshi Yoshii, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G Okuno. Improving Efficiency and Scalability of Model-Based Music Recommender System Based on Incremental Training. In *ISMIR 2007: proceedings of the 8th International Conference of Music Information Retrieval*, number ISMIR, 2007.

A spectral clustering method for musical motifs classification

Alberto Pinto

Quintade Research
Viale San Gimignano, 4
20146 Milano (Italy)
apinto@quintade.org
apinto@ccrma.stanford.edu

Abstract. In recent years, spectral clustering methods are getting more and more attention in many fields of investigation for analysis and classification tasks. Nevertheless, no applications to symbolic music have been provided yet.
Here we present a method for motif classification based on spectral clustering of music scores that can be exploited, for instance, in automatic or computer-assisted music analysis. Scores are represented through a network-graph of segments and then ranked depending on their centrality within the network itself, which can be measured through the components of the leading eigenvector associated to the Laplacian of the graph. Moreover, segments with higher centrality are more likely to be relevant for music summarization.
An experimental musicological analysis has been performed on J.S.Bach's 2-part Inventions to prove the effectiveness of the method.

Keywords: spectral clustering, graph, centrality

1 Introduction

The problem of automatically identifying relevant characteristic motifs and efficiently store and retrieve the digital content has become an important issue as digital collections are increasing in number and size more or less everywhere. Music segmentation is usually realized through musicological analysis by human experts and, at the moment, automatic segmentation is a difficult task without human intervention. The supposed music themes have often to undergo a handmade musicological evaluation, aimed at recognizing their expected relevance and completeness of results. As a matter of fact, an automatic process could extract a musical theme which is too long, or too short, or simply irrelevant. That is why a human feedback is still required in order to obtain high-quality results.

Some proposed automatic methods are more focused on tonal music as they exploit the harmonic structures of a piece and voice leading. On the other hand, other methods are more general and do not take into account neither harmony nor rhythm.

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

Notwithstanding the conspicuousness of the literature, current approaches seem to rely just on repetitions [1] [2] [3], assigning higher scores to recurring equivalent melodic and harmonic patterns [4]. Recently reported approaches to melodic clustering based on motivic topologies [5], graph distance [6] [7] and paradigmatic analysis [8] have been used to select relevant subsequences among highly repeated ones by heuristic criteria [9] [10].

Moreover, the "paradigm of repetition", in order to be applied, needs by no means a precise definition of "varied repetition", a concept not easy to define. Of course, it has to include standard music transformation, but it is very difficult to adopt a simple two-valued logic (this is a repetition and this is not) in this context, where a more fuzzy approach seems to better address such a problem.

Here we present a ranking method based on relations instead of repetitions. We show that a distance distribution on a graph of note subsequences induced by music similarity measures generates a ranking real eigenvector whose components reflect the actual relevance of motives. Spectral ranking on this eigenvector allows to better identify different sections within a piece through the partitioning of the score into clusters of similar melodies.

2 Related approaches

Lartillot [11] [12] defined a musical pattern discovery system motivated by human listening strategies. Pitch intervals are used together with duration ratios to recognize identical or similar note pairs, which in turn are combined to construct similar patterns. Pattern selection is guided by paradigmatic aspects and overlaps of segments are allowed.

Cambouropoulos [13], on the other hand, proposed methods to divide given musical pieces into mostly non-overlapping segments. A prominence value is calculated for each melody based on the number of exact occurrences of non-overlapping melodies. Prominence values of melodies are used to determine the boundaries of the segments [14]. He also developed methods to recognize variations of filling and thinning (through note insertion and deletion) into the original melody. Cambouropoulos and Widmer [15] proposed methods to construct melodic clusters depending on the melodic and rhythmic features of the given segments. Basically, similarities of these features up to a particular threshold are used to determine the clusters. High computational costs of this method make applications to long pieces difficult.

2.1 Tonal harmony-based approaches

Tonal harmony based approaches exploit particular harmonic patterns (such as tonic-subdominant-dominant-tonic), melodic movements (e.g. sensible-tonic), and some rhythmical punctuation features (pauses, long-duration notes, ...) for a definition of a commonly accepted semantic in many ages and cultures.

These approaches typically lead towards score reductions (see Figure 1), made possible by taking advantage of additional musicological information related to

the piece and assigning different level of relevance to the notes of a melody. For example one may choose to assign higher importance to the stressed notes inside a bar [16]. In other words, the goal of comparing two melodic sequences is achieved by reducing musical information into some "primitive types" and comparing the reduced fragments by means of suitable metrics.

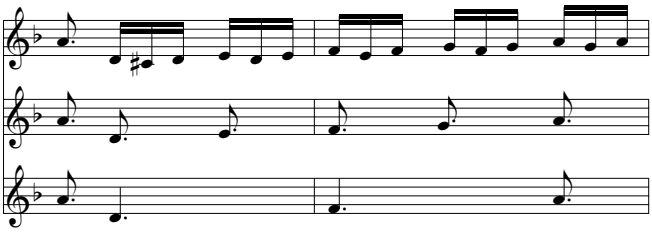


Fig. 1. J.S. Bach, BWV 1080: Score reductions.

A very interesting reductionistic approach to music analysis has been attempted by Fred Lerdahl and Ray Jackendoff. Lerdahl and Jakendoff [17] research was oriented towards a formal description of the musical intuitions of a listener who is experienced in a musical idiom. Their purpose was the development of a formal grammar which could be used to analyze any tonal composition.

The study of these mechanisms allows the construction of a grammar able to describe the fundamental rules followed by human mind in the recognition of the underlying structures of a musical piece.

2.2 Topological approaches

Mazzola and Buteau [18] proposed a general theoretical framework for the paradigmatic analysis of the melodic structures. The main idea is that a paradigmatic approach can be turned into a topological approach. They consider not only consecutive tone sequences, but allow any subset of the ambient melody to carry a melodic shape (such as rigid shape, diastematic shape, etc.). The mathematical construction is very complex and, as for the motif selection process, it relies on the repetition paradigm.

The method proposed by Adiloglu, Noll and Obermayer in [10] does not take into account the harmonic structure of a piece and is based just on similarities of melodies and on the concept of similarity neighborhood. Melodies are considered as pure pitch sequences, excluding rests and rhythmical information.

A monophonic piece is considered to be a single melody M, i.e. they reduce the piece to its melodic surface. Similarly, a polyphonic piece is considered to be the list $M = (M_i)_{i=1,\dots,N}$ of its voices M_i. The next step is to model a number of different melodic transformations, such as transpositions, inversions and retrogradations and to provide an effective similarity measure based on

cross-correlation between melodic fragments that takes into account these transformations. They utilize a mathematical distance measure to recognize melodic similarity and the equivalence classes that makes use of the concept of *neighbourhood* to define a set of similar melodies.

Following the repetition paradigm stated by Cambouropoulos in [14] they define a prominence value to each melody based on the number of occurrences, and on the length of the melody. The only difference is that they allow also melody overlapping. In the end, the significance of a melody m of length n within a given piece M is the normalized cardinality of the similarity neighbourhood set of the given melody. If two melodies appear equal number of times, the longer melody is more significant than the shorter one.

In [10] the complete collection of the Two-part Inventions by J. S. Bach is used to evaluate the method, and this will be also our choice in section 4.

3 The model

Our point of view can be synthesized in the following points:

1. consider a music piece as a network graph of segments,
2. take into account both melodic and rhythmical structures of segments
3. do not consider harmony, as it is too much related to tonality.

A single frame may represent, for instance, a bar or a specific voice within a bar like in Fig. 2, but also more general segments of the piece. Thus, a music piece can be looked at like a complete graph K_n. In graph theory, a complete graph is a simple graph where an edge connects every pair of distinct vertices. The complete graph on n vertices has $n(n-1)/2$ edges and is a regular graph of degree $n-1$. In this representation, score segments correspond to graph nodes and the similarity between couples of segments correspond to edge weights.

3.1 Metric weights

In this Section we are going to introduce the metric concepts we adopted to calculate similarities between different score windows. The variety of segmentations reflects to a large extent the variety of musical similarity concepts, nevertheless, as stated in Section 4, the model is rather robust respect to metric changes.

In general, we can just require that the set of segments can be endowed with a notion of distance

$$d : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$$

between pairs of segments and turns this set into a (possibly metric) space $(\mathcal{S}, d)$. A natural choice for point sets of a metric space is the Hausdorff metric [19] but any other distance discovered to be useful in music perception, like EMD/PTD [20], can be assumed as well.

Here we assume d to be:

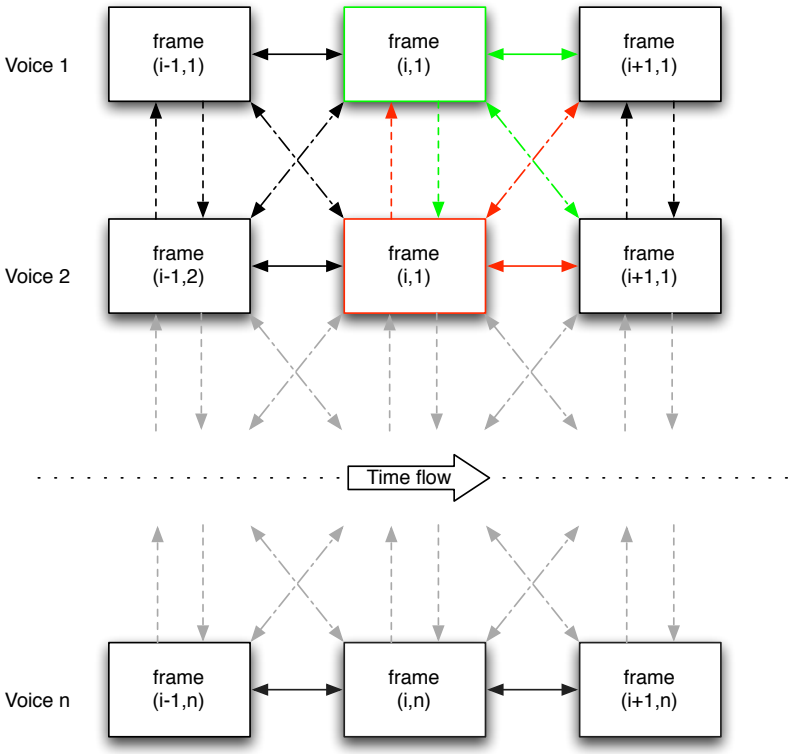


Fig. 2. A representation of the (first-order) network of frames.

1. real,
2. non-negative,
3. symmetric and
4. such that $d(s, s) = 0$, $\forall s \in \mathcal{S}$

As a matter of fact, most musically relevant perceptual distances do not satisfy all metric axioms [20]. Therefore no further property, like the identity of indiscernibles or the triangle inequality, is assumed.

Given two segments s_1 and s_2, the metrics we adopted in the experiments are the following:

$$d_1(s_1, s_2) = \sqrt{\sum_{|s|} |[s_1]_{12} - [s_2]_{12}|^2} \tag{1}$$

$$d_2(s_1, s_2) = \sqrt{\sum_{|s|} (s_1'(t) - s_2'(t))^2} \tag{2}$$

where s' is the derivative operator on the sequence s, $|s|$ is the length of s and $[s]_{12}$ is the sequence s where each entry has been chosen in the interval $[0, 11]$.

d_1 is a first-order metric that takes into account just octave transpositions of melodies. In fact, pitch classes out of the range $[0, 11]$ are folded back into the same interval, so melodies which differ for one or more octaves belong to the same congruence class modulo 12 semitones. d_2 is a second-order metric that takes into account arbitrary transpositions of a melody. No other assumptions on possible variations have been made, so that an equivalence class of melodies is composed just of transpositions and inversions of the same melody like in Adiloglu (2006).

Both distances can be applied to single voice sequences but also to multiple voice sequences, given that a suitable representation has been provided. For instance, in a two voice piece, with voices v_1 and v_2, one can consider the difference vector $v = v_1 - v_2$ as a good representation of a specific segment, and then apply d_1 or d_2 to this new object. The advantage of using this differential representation is that it is invariant respect to transpositions of the two voices so that, for instance, it makes also d_1 invariant respect to transpositions, and not just to octave shifts.

By exploiting those distance concepts, it is possible to endow the edges of the complete graph with metric weights in order to compute the weights of nodes in terms of the main eigenvector, as we are going to show in the following Sections.

3.2 The algorithm

Let $d : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ denote a distance function on $\mathcal{S}$, like those defined in Section 3.1, which assigns each pair of segments s_i and s_j a distance $d(s_i, s_j)$. We can describe the algorithm through the following steps:

1. Form the distance matrix $A = [a_{i,j}]$ such that $a_{i,j} = d(s_i, s_j)$;
2. Form the affinity matrix $W = [w_{i,j}]$ defined by

$$w_{i,j} = \exp(-\frac{-d^2(x_i, x_j)}{2\sigma^2}) \tag{3}$$

   The parameter σ can be chosen experimentally, a possible choice is the standard deviation of the similarity values within the considered network graph (this has been our choice in the experimental part);
3. Form the Laplacian matrix $L = D^{-1/2}WD^{-1/2}$, where D is the diagonal matrix whose (i, i) element is the sum of W's i-th row
4. Compute the leading eigenvector $x = [x_i]$ of L and rank each segment s_i according to the component x_i of x.
5. Perform a k-means algorithm on the leading eigenvector to cluster the segments.

4 Experimental results

In order to evaluate the relevance of the results of the proposed method we need a suitable data collection together with a commonly acceptable ground truth

for that collection. Following [10], Johann Sebastian Bach's *Two-part Inventions* has been our choice. For this collection, a complete ground truth is provided by musicological analysis and it can be found for example in [21] and [22].

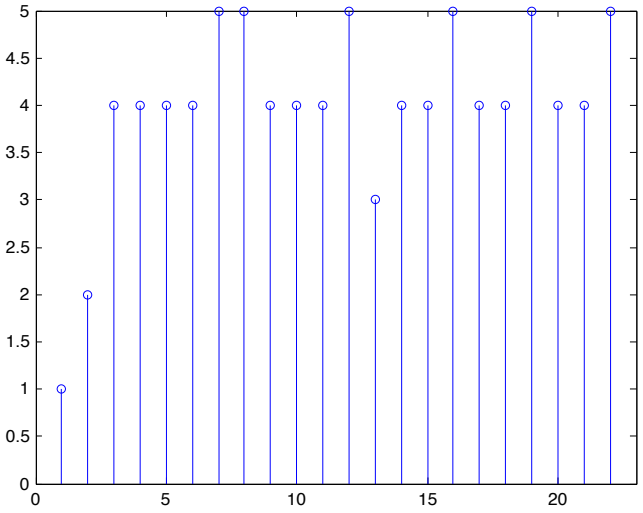


Fig. 3. Clustered bars in BWV 772 according to k-means performed on the leading eigenvector of the laplacian matrix.

When compared to musicological analysis [10] [21] [22] it is evident that the centrality-based model outperforms the repetition-based model, providing also more significative information. Segments with higher rank in the relational model represent always relevant bars of the score, even if they may be different by using different metrics. This means that relevant bars contain a main motif or characterizing sequences. It is not the same for the model based on repetitions: here the relevancy really depends just on the number of repetitions, so it can happen that a trill turns to be more relevant than the rest of the piece just because its repetition rate is higher than that of the other bars.

Model	Precision (%)
Repetition	43
d_1	77
d_2	95

Table 1. Precision results for the three models applied to J. S. Bach's Inventions.

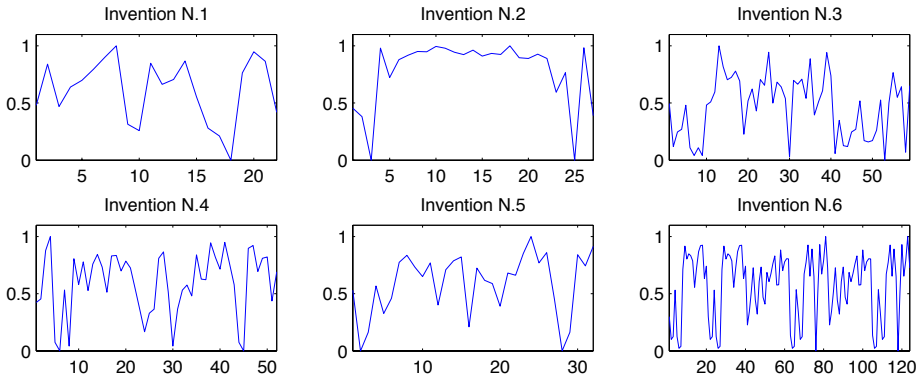


Fig. 4. Centrality values plotted against bar numbers for the first 6 J.S.Bach's Two-Part Inventions.

Bar ranking is in principle not affected by the repetition rate of patterns and higher importance is equally given to higher and lower repetition rates. Of course, superpositions of the two methods may happen too.

On the other hand, cases exist for which no repetition occurs and, consequently, the repetition paradigm is not applicable in principle, unless defining ad hoc neighborhood concepts for each piece. In these cases, motif centrality can provide significant results.

In Figures 4 and 5 the components of the main eigenvector for each invention, representing the degree of centrality of each bar within the network graph, have been plotted against bar numbers. This provides an immediate representation of the importance of each bar within the whole piece. Bars with higher values are more likely to contain a main motif of the piece.

Figure 3 reports the results for bar spectral clustering in the case of BWV 772 according to k-means, with k=5, performed on the leading eigenvector of the laplacian matrix. It is evident how the main theme which appears in the first two bars is identified in the first two clusters.

5 Conclusions

We presented an approach for motif discovery in music pieces based on an eigenvector method. Scores are segmented into a network of bars and then ranked depending on their graph centrality. Spectral is performed in order to classify all the bar segments. Bars with higher centrality grouped into the same cluster can be exploited for music summarization. Experiments performed on the collection of J.S.Bach's 2-parts Inventions show the effectiveness of the method.

Further investigations deal, for instance, with the relationships between particular mathematical entities (e.g. spectra) and particular musical issues (e.g. genre, authorship).

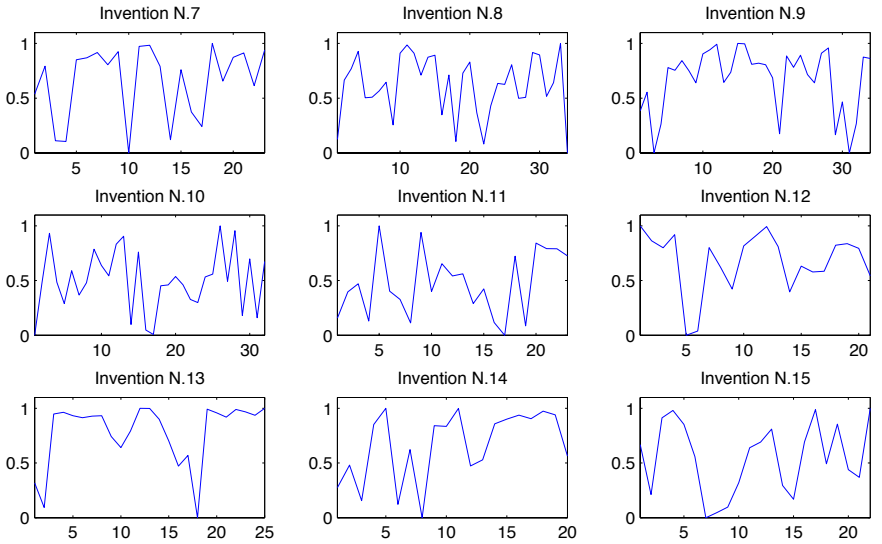


Fig. 5. Centrality values plotted against bar numbers for the last 9 J.S.Bach's Two-Part Inventions.

Second, one could investigate how different metrics d relate to different concepts of melodic and harmonic similarity and how this is related to cluster stability. In this context, the inverse problem of finding metrics d induced by a priori eigenvectors (coming from a hand-made musicological analysis) could provide interesting insights into music similarity perception.

Finally, it is also possible to compare different music pieces from a structural point of view by comparing their segmentation derived from spectral clustering.

References

1. Pienimaki, A.: Indexing Music Databases Using Automatic Extraction of Frequent Phrases. Proceedings of the International Conference on Music Information Retrieval (2002) 25–30
2. Cambouropoulos, E., Crochemore, M., Iliopoulos, C., Mouchard, L., Pinzon, Y.: Algorithms for computing approximate repetitions in musical sequences. International Journal of Computer Mathematics **79**(11) (2002) 1135–1148
3. Livingstone, S., Palmer, C., Schubert, E.: Emotional response to musical repetition. (2011)
4. Crawford, T., Iliopoulos, C., Raman, R.: String Matching Techniques for Musical Similarity and Melodic Recognition. Computing in Musicology **11** (1998) 73–100
5. Mazzola, G., Müller, S.: The Topos of Music: Geometric Logic of Concepts, Theory, and Performance. Birkhäuser (2002)
6. Pinto, A.: Mining music graphs through immanantal polynomials. In: In Proceedings of the 6th International Workshop on Mining and Learning with Graphs. (2008)

7. Pinto, A.: Multi-model music content description and retrieval using IEEE 1599 XML standard. Journal of Multimedia **4**(1) (2009) 30
8. Nestke, A.: Paradigmatic Motivic Analysis. Perspectives in Mathematical and Computational Music Theory, Osnabrück Series on Music and Computation (2004) 343–365
9. Lartillot, O., Saint-James, E.: Automating Motivic Analysis through the Application of Perceptual Rules. Music Query: Methods, Strategies, and User Studies (Computing in Musicology) **13** (2004)
10. Adiloglu, K., Noll, T., Obermayer, K.: A paradigmatic approach to extract the melodic structure of a musical piece. Journal of New Music Research **35**(3) (2006) 221–236
11. Lartillot, O.: Discovering musical patterns through perceptive heuristics. Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR 2003) (2003) 89–96
12. Lartillot, O.: A musical pattern discovery system founded on a modeling of listening strategies. Comput. Music J. **28**(3) (2004) 53–67
13. Cambouropoulos, E.: Extracting'Significant'Patterns from Musical Strings: Some Interesting Problems. Presente aux London String Days **2000** (2000)
14. Cambouropoulos, E.: Musical pattern extraction for melodic segmentation. Proceedings of the ESCOM Conference 2003 (2003)
15. Cambouropoulos, E., Widmer, G.: Automated motivic analysis via melodic clustering. Journal of New Music Research **29**(4) (2000) 303–318
16. Selfridge-Field, E.: Towards a Measure of Cognitive Distance in Melodic Similarity. Computing in Musicology **13** (2004) 93–111
17. Lerdahl, F., Jackendoff, R.: A Generative Theory of Tonal Music. MIT Press, Cambridge, Massachusetts (1996)
18. Buteau, C., Mazzola, G.: From Contour Similarity to Motivic Topologies. Musicae Scientiae **4**(2) (2000) 125–149
19. Di Lorenzo, P., Di Maio, G.: The Hausdorff Metric in the Melody Space: A New Approach to Melodic Similarity. In: Ninth International Conference on Music Perception and Cognition. (2006)
20. Typke, R., Wiering, F., Veltkamp., R.C.: Transportation distances and human perception of melodic similarity. Musicae Scientiae, Discussion Forum 4A, 2007 (special issue on similarity perception in listening to music), p. 153-182.
21. Derr, E.: The Two-Part Inventions: Bach's Composers' Vademecum. Music Theory Spectrum **3** (1981) 26–48
22. Williams, P.: JS Bach. Cambridge University Press

Songs2See: Towards a New Generation of Music Performance Games

Estefanía Cano, Sascha Grollmisch, and Christian Dittmar *

Fraunhofer Institute for Digital Media Technology
Ilmenau, Germany
{cano,goh,dmr}@idmt.fraunhofer.de

Abstract. In this paper, the concept of *Music Performance Games* is introduced and contrasted with related terms like music video games, interactive applications, and serious games. The game Songs2See is introduced as an example of *Music Performance Games* and its design stage is evaluated within a conceptual framework for serious game development. Future directions for improvement and testing of the game are outlined.

Keywords: Music game, education, score, rhythm, pitch, feedback, performance

1 Introduction & Related Work

Music video games can be defined as games where the gameplay is mostly oriented to the user's interaction with music, scores, songs or music performance. In general, the term music video game encompasses a wide variety of game categories such as dancing games, interactive composition games, rhythm games, pitch games, music management games, etc. The earliest music games developed were rhythm games where the user is required to follow a sequence or pattern of instructions such as pressing different buttons on a game controller. Popular examples of rhythm games are Guitar Hero[1] and Rock Band[2]. Later on and due to the development of solid pitch detection algorithms, pitch games became popular. In these games, the user's ability to match the pitch of a piece of music is tested. There is a clear rhythmic element in pitch games—as users are requested to produce a certain pitch at a specific time; however, the main focus of the game is on the correct intonation of a series of notes. A popular example of a pitch game is the karaoke game Singstar[3].

* The Thuringian Ministry of Economy, Employment and Technology supported this research by granting funds of the European Fund for Regional Development to the project Songs2See, enabling transnational cooperation between Thuringian companies and their partners from other European regions.

[1] Guitar Hero: http://www.guitarhero.com

[2] Rock Band: http://www.rockband.com

[3] Singstar: http://www.singstar.com

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

From an educational point of view, music video games play an important role in creating interest in music performance and musical instruments. However, transferring the skills developed in the game to the performance of real musical instruments, is not a straight-forward process [7]. For obvious reasons, game controllers cannot capture the real characteristics and intricacies of a musical instrument and are in general extremely simplified versions of them.

In the rapidly changing and technological environment where new generations grow and learn, educational methods needed to evolve correspondingly to fit their needs and life styles [3]. For this reason, interactive applications for music learning have also been developed. Here, the idea is to take advantage of the various possibilities provided by digital audio, video, and software developments to design learning applications that can support educational processes. Some commercial applications for music learning are Music Delta[4] and Smart Music[5]. Currently, the project Kopra-M[6] deals with measurement of competencies in music. For this matter, a systematic methodology and a proprietary software solution to assign and control music tasks is developed. The outcomes of this project are targeted to German secondary school students.

In the past years a few research projects have dealt with the development of E-learning systems for music education. The IMUTUS[7] (Interactive Music Tuition System), the VEMUS[8] (Virtual European Music School), and the i-Maestro[9] (Interactive Multimedia Environment for Technology Enhanced Music Education and Creative Collaborative Composition and Performance), were all European based projects partially funded by the European Commission that addressed music education from an interactive point of view. See [4] for a thorough description of these projects.

As a meeting point between video games and interactive learning applications, *Serious Games* have been developed. *Serious games* have been defined as entertaining games with non-entertainment goals. They educate, train, inform, and aim at the achievement of a predefined objective through a gaming experience [3]. Games offer an ideal medium for introducing new skills and knowledge [5] and provoke active learner involvement through exploration, experimentation, competition and co-operation [3]. They support learning because of increased visualization and challenged creativity. *Serious games* have been developed for different scenarios such as: military, humanitarian, social, business, commercial, etc. FloodSim[10], ShipSim[11], NanoMisson[12], and Food Force[13] are all examples

[4] Music Delta: `http://www.griegmusic.dreamhosters.com/?page_id=98`
[5] Smart Music: `http://www.smartmusic.com`
[6] Kopra-M: `http://www.idmt.fraunhofer.de/de/projekte/laufende_projekte/komus.html`
[7] IMUTUS: `http://www.exodus.gr/imutus/index.htm`
[8] VEMUS: `http://www.tehne.ro/projects/vemus_virtual_music_school.html`
[9] i-maestro: `http://www.i-maestro.org/`
[10] FloodSim: `http://www.floodsim.com/`
[11] ShipSim: `http://www.shipsim.com/`
[12] NanoMission: `http://nanomission.org/`
[13] Food Force: `http://www.wfp.org/how-to-help/individuals/food-force`

of serious games. In the music field, games like Rocksmith[14], and Wild Chords[15] present structured learning goals to be achieved through the game.

In the context of this paper, the term *music performance games* will be used to refer to music games dealing with performance aspects of music such as musical instruments, rhythm and pitch. Three important aspects are considered within the definition of *music performance games*: (1) As for all games, entertainment and immersion should be critical elements. (2) The game must directly involve the production of musical sound. In other words, the actions performed by the user during the game sequence should directly result in the production of musical sounds— singing, playing a musical instrument, synthesizing sound, etc. (3) The game should attempt to achieve a specific goal within the performance aspects of music. Common examples are: playing a selected tune on the trumpet, playing major scales fluidly, singing different intervals in tune or learning to play specific chords on the guitar.

The remainder of this paper is organized as follows: Section 2 describes the game Songs2See and its main features, Section 3 introduces a conceptual framework for serious game development, Section 4 outlines some final remarks. Finally, conclusions are presented in Sections 5.

2 Songs2See

2.1 General Overview

Songs2See[16] is a *music performance game* where users can practice a selected musical piece on their own musical instrument. The basic concept behind the game is that users play to the computer microphone and the system evaluates their performance in real-time. The Songs2See Game is complemented by the Sons2See Editor. This is a software application that allows user to create their own musical exercise content from mp3 or wav files. Both in the Songs2See Game and in the Songs2See Editor, state-of-the-art music information Retrieval (MIR) techniques for pitch detection, sound separation, music transcription, and beat extraction have been applied [1, 2, 4, 6]. Figure 1 shows the game interface. It is composed of two main elements: (1) The Game View, (2) The Instrument View.

The Game View: In this part of the interface, the sequence of notes of the chosen melody is displayed. The idea behind the design of the Game View was to include as many musical elements as possible without requiring the user to have previous musical knowledge or music reading skills. The upper-half of Figure 1 shows the Game View. Some details to note:

1. A complete musical staff is displayed. The following elements were considered:

[14] Rocksmith: `http://www.rocksmith.com`
[15] WildChords: `http://www.wildchords.com/`
[16] Songs2See: `http://www.idmt.fraunhofer.de/en/Service_Offerings/technologies/q_t/songs2see.html`

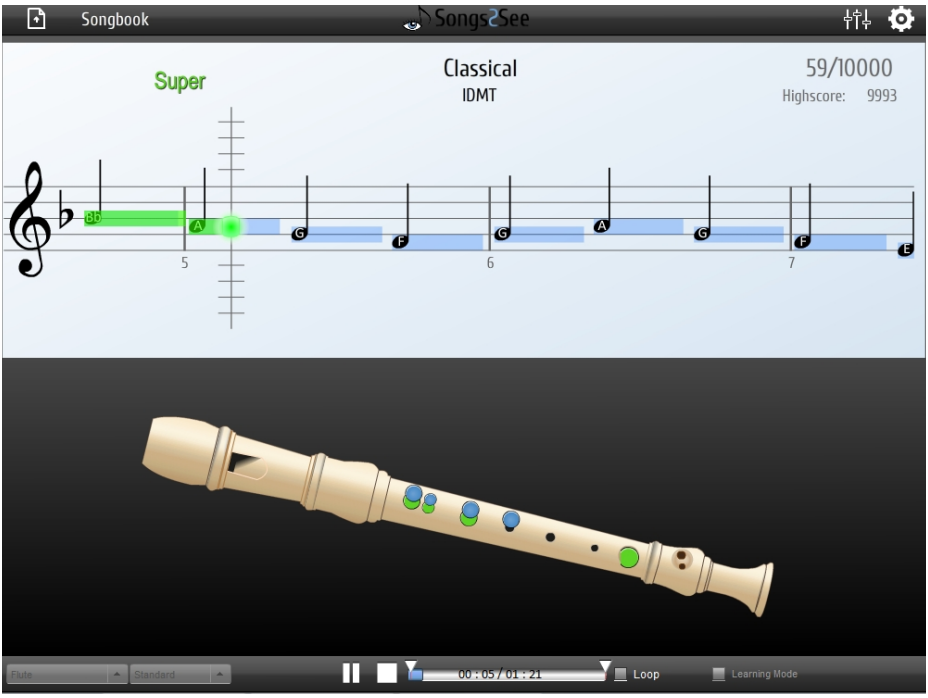


Fig. 1: The game interface. (1) Game View (upper-half) (2) Instrument View (lower-half)

- Pitch: displayed both by placing the elements in their corresponding locations in the staff and by displaying note names inside the note heads.
- Rhythm: displayed using normal music notation and length bars in the scrolling score. The time signature (see Figure 2) and bars are also displayed to give time references to the user.
- Key: the flat and sharp signs are displayed as in normal scores and key reminders are shown in the note names. In Figure 1, the key of the song is D minor equivalent to one flat (Bb). The name of the first note shown in the sequence is then Bb.

2. Real-time feedback is given in different ways: (1) The note played by the user is displayed at all times by the *Pitch Marker*. (2) When the correct note is played, the length bars are colored in green. (3) Three different colored signs are displayed to guide the user—Super, OK and Missed in green, yellow and red respectively. In Figure 1 a green "Super" sign is displayed. (4) The score obtained by the user and the highest personal score are displayed.

The Instrument View: The game supports the use of piano, bass, guitar, saxophone, trumpet, flute and voice. In the *Instrument View*, the game presents an

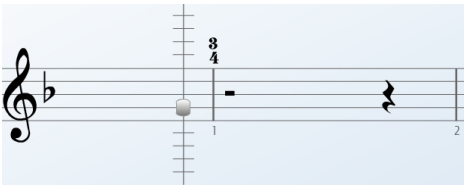
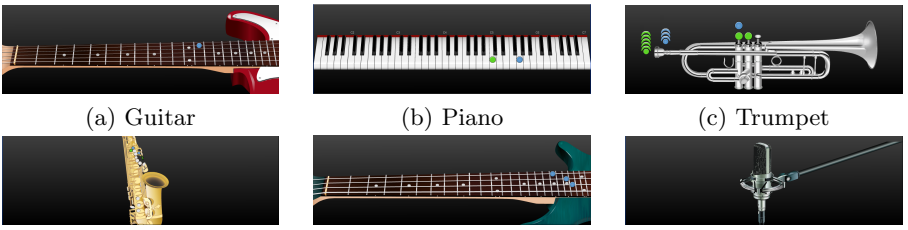


Fig. 2: Time signature displayed in the Game View.

automatic fingering animation that guides the user through the melody sequence of the chosen musical piece. For all the instruments supported, the fingering of the current note is displayed in green and the fingering of the next note in the sequence is displayed in blue. This intends to guide the user in the transition between fingering positions.



(a) Guitar (b) Piano (c) Trumpet

(d) Saxophone (e) Bass (f) Voice

Fig. 3: Instruments supported. The flute is displayed in Figure 1

3 A Conceptual Framework for Serious Game Design

In [8], a conceptual framework for serious game development has been presented. The main goal of this framework is to provide game designers and educators with a conceptual model to guide the development of serious games to be effective in the achievement of the learning goals. This model is composed of nine different elements shown in Figure 4, which will be briefly explained. Furthermore, the Songs2See game will be revised in the context of this framework.

1. Capability: Refers to the cognitive, psychomotor and affective skills that the user is to develop in the game.
   In Songs2See, the intended musical capabilities to develop are: fluid performance, effective identification, consistent element relation, timely execution, and thorough understanding.
2. Instructional Content: Refers to the facts, procedures, concepts, and principles that users should learn.

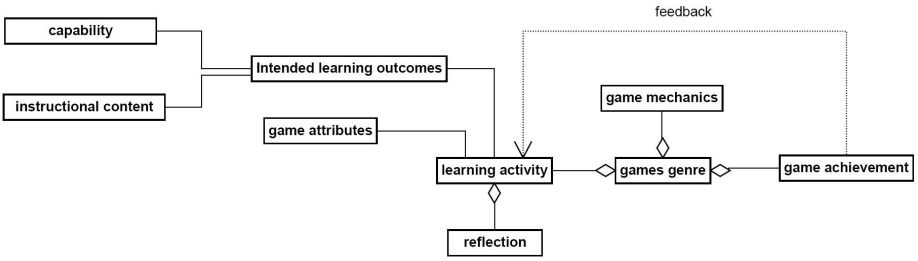


Fig. 4: Block diagram of the conceptual framework proposed by Yussof et al.

It is intended that the user understands and learns concepts as melody, notes, rhythm, fingerings, and instrumental basics with the use of Songs2See.

3. Intended learning outcomes: Can be seen as a combination of capability and instructional content. They refer to the goals to be achieved from playing the game.
   In Songs2See the intended learning outcomes are: fluid performance of a melody or musical piece, effective identification of notes and of their corresponding pitches, timely execution of a sequence of notes with their corresponding durations, rapid identification of fingerings, consistent relation of notes with their corresponding fingerings, understanding of instrument mechanics and sound production principles.
4. Game attributes: These are the elements of the game that support learning and engagement.
   In Songs2See several elements have been included with the goal of making the learning process more entertaining and suitable for all users [4]. *Incremental learning* for example, is supported by the included Learning Mode. Users can practice new pieces step by step until they are confident enough to play them at normal speed. *Instructional scaffolding*[17] is supported by different elements: the automatic fingering animation helps users relate pitches to fingering positions in the instrument, the option of including note names in the score helps students in the process of learning standard music notation. Additionally, rhythm learning is supported by the length bars presented in the scrolling score. *Interaction* is supported as the user is constantly presented with melodies that require timely responses and actions to be performed correctly. Furthermore, *Learner control*[18] is supported by the possibility of loading music pieces that fit the user's taste and skills. *Feedback* is given to the user in real-time so there is continuous awareness of the outcome of the performance. At the end of each performance, *rewards* are given in the form of scores based on a rating system. With the use of real musical instruments and with the inclusion of elements to bring the game close to a

[17] Refers to the support given to the user to promote learning of new concepts
[18] Possibility given to the user to direct their learning experience to fit their own pace and progress

real performance scenario (e.g. real-time performance conditions, accompaniment tracks, music notation elements), an attempt is made to make the process of transferring skills and experiences of the game to the real world, as smooth as possible. This concept is known as *authentic learning*.

5. Learning activity: Refers to the activities designed to provide engagement an immersion in the game.
   In Songs2See two different types of learning activities are included: (1) Performance of music pieces selected by the user in the instrument of their choice. (2) Practice sets with specially designed content to address topics as scales, chord or intervals.
6. Reflection: Is the process where the user is given the opportunity to think about the learning goal and reflect about the strategy to be taken in the next activity.
   In Songs2See, this is presented in real-time. When the user plays the wrong note, the game displays the note played. The intention is that users can contrast their performances with the correct melody and possibly understand the cause of error.
7. Games genre: Refers to type or category of the game to be played. As defined in Section 1, Songs2See can be classified as a *music performance game*.
8. Game mechanics: Conditions and rules that define the details of the game.
   In Songs2See the game mechanics are simple. The user is to perform the selected piece of music with the chosen instrument in the attempt to timely and fluidly follow the sequence of notes presented by the game.
9. Game achievement: Refers to the user's level of achievement in the game.
   In Songs2See a final score is given to the user after each performance. Furthermore, a record of high scores is kept in the game to give the user some insight of previous achievements. An additional option that relates to game achievement is the tolerance value in the settings of the game. The user can select how strict the rating system should be when evaluating the performance and consequently be challenged to more accurate performances.

4 Final Remarks

The importance of analyzing Songs2See within this conceptual framework lies on the fact that clear pointers on how to improve the game and its learning potential can be obtained. In terms of *reflection* for example, offering the user the possibility of reviewing performances off-line, can potentially improve the understanding of the source of error. However, the challenge lies on doing it without withdrawing the user from the gaming environment. A plausible solution could be the design of mini-games as part of the *Learning Activities*. Here, only segments with clear difficulties— most likely selected by the users themselves— can be addressed. The mistakes made in the initial performance can be highlighted so the user can be aware of the problematic passages. Another possibility to increase *reflection* in the game, could be incorporating a recording option. Users could playback an animated version of the performance, where

mistakes are highlighted for reference. This animation could be paused and replayed as many times as the user finds it necessary. In terms of possible *learning activities*, specially designed content could be created to familiarize users with instrument mechanics. Animated description of the instruments, instructions on how to hold them, and basic explanations on sound production are all possibilities within the game. A reward strategy could also be devised to support *game achievements*. This could increase users' interest in outperforming themselves. A possibility could be offering sets of content or exercises, where different levels can only be reached after achieving a certain score in previous levels.

5 Conclusions

The concept of *Music Performance Games* has been presented and as an example, the game Songs2See has been described. As a guide through the development stage of Songs2See, the game has been evaluated within the conceptual framework for serious game development presented in Section 3. Several pointers on how to improve the design of the game and its learning potential were presented. The main goal of placing Songs2See within this framework was to optimize the designing stage so the final outcome and potential of the game are also maximized. However, the learning capabilities, engagement and effectiveness of the game can only be measured when it is delivered to the final user.

References

1. E. Cano, C. Dittmar, and S. Grollmisch. Songs2See : Learn to Play by Playing. In *12th International Society for Music Information Retrieval Conference (ISMIR)*, Miami, USA, 2011.
2. E. Cano, C. Dittmar, and G. Schuller. Efficient Implementation of a System for Solo and Accompaniment Separation in Polyphonic Music. In *20th European Signal Processing Conference*, Bucharest, Romania. Submitted.
3. J. c.k.h Riedel and J. B. Hauge. State of the Art of Serious Games for Business and Industry. In *17th International Conference on Concurrent Enterprising (ICE 2011)*, —Aachen, Germany, 2011.
4. C. Dittmar, E. Cano, and J. Abeß er. Music Information Retrieval Meets Music Education. *Dagstuhl Follow-Ups: Multimodal Music Processing*, pages 94–117, 2012. To appear.
5. T. M. Doll, R. Migneco, and Y. E. Kim. Web-based Sound and Music Games with Activities for STEM Education. *2009 International IEEE Consumer Electronics Society's Games Innovations Conference*, pages 191–200, Aug. 2009.
6. S. Grollmisch, C. Dittmar, E. Cano, and K. Dressler. Server based pitch detection for web applications. In *AES 41st International Conference: Audio for Games*, London, UK, 2011.
7. S. Grollmisch, C. Dittmar, and G. Gatzsche. Concept, Implementation and Evaluation of an improvisation based music video game. *Proceedings of IEEE Consumer Electronics Society's Games Innovation Conference (IEEE GIC)*, 2009.
8. A. Yusoff, R. Crowder, L. Gilbert, and G. Wills. A Conceptual Framework for Serious Games. *Ninth IEEE International Conference on Advanced Learning Technologies*, pages 21–23, July 2009.

A Music Similarity Function Based on the Fisher Kernels

Jin S. Seo[1], Nocheol Park[1], and Seungjae Lee[2] *

[1] Dept. of EE, Gangneung-Wonju National University, Korea
[2] Creative Content Research Laboratory, Electronics and Telecommunications Research Institute, Daejeon, Korea
jsseo@gwnu.ac.kr, seungjlee@etri.re.kr

Abstract. Music-similarity computation is an essential building block for browsing, retrieval, and indexing of digital music archives. This paper presents a music similarity function based on the Fisher-vector representation of the spectral features extracted from a song. The distance between the Fisher vectors of two songs is used as the similarity of the two songs. The Fisher vector has a closed-form representation and can be readily incorporated with simple vector distance measures. Experimental results show that the Fisher-vector representation of the auditory features is promising for the music-similarity computation.

Keywords: music similarity, music retrieval, music browsing, Fisher kernel

1 Introduction

Computing similarity between two songs is essential for browsing, retrieval, and indexing of digital music archives. Music similarity can be inferred in two different ways; collaborative filtering and content-based approach. In collaborative filtering, based on the musical tastes of many people, the musical preference of one person is predicted by those of other people [1]. In content-based approach, based on the perceptual auditory features, music similarity is directly computed from the distance between features from two songs. Both approaches have pros and cons. For example, the collaborative filtering cannot be adopted for new songs, and the content-based approach requires perceptually-meaningful feature extraction and computationally-efficient distance measure. This paper deals with the content-based approach.

The difficulty in computing the music similarity lies in the fact that the criteria used to determine the level of the similarity between two songs are subjective and hard to be described quantitatively. For the content-based music similarity, auditory features representing the music timbre, such as mel-frequency cepstral

* The authors would like to thank Seokjeong Lee for insightful discussions and help in collecting a dataset. This research project was supported by Government Fund from Korea Copyright Commission.

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

coefficients (MFCC) or other spectrum descriptors, has been adopted. In [1][2], the low-level spectral features extracted from a song are modeled by the k-means cluster or Gaussian Mixture Model (GMM). The distance between the song-level representations is estimated by either KL divergence [2][3], or earth-mover distance (EMD) [1], which is used as a metric for music similarity. Despite their excellent performance, the above mentioned methods are characterized by several short-comings. First of all, the construction of the song-level representations is based on an iterative process, which may not converge in some cases. Second, the pairwise distance using the KL or the EMD is computationally expensive and does not have a closed-form solution in most of the cases. To mitigate these problems, we employ the Fisher kernel to represent the feature distribution of a song instead of the iterative modeling process. The Fisher kernel was first introduced by Jaakkola and Haussler [4] and further studied by Perronnin and Dance [5] for image classification [6] and retrieval [7]. To combine the benefits of generative and discriminative approaches, the key idea of the Fisher kernel is to characterize a signal with a gradient vector derived from a probability density function which models the generation process of the signal [5][6][7]. In this paper, the closed-form vector representation of the Fisher kernel (the Fisher vector) derived in [5] is applied to represent the auditory features of the songs and compared with each other using simple distance measures, such as the Euclidean or the Cosine distance. In the experiments, the music similarity function based on the Fisher-vector representation showed retrieval performance comparable to the previous one [1].

This paper is organized as follows. Section 2 describes the music-similarity computation based on the Fisher-vector representation. Section 3 presents the experimental results of music retrieval tests. Finally, section 4 summarizes the paper.

2 Music Similarity Based on the Fisher-Vector Representation

The overview of the content-based music similarity computation is shown in Fig. 1. In the previous methods [1][2], the underlying distribution of the spectral features from a music clip is used as a signature for the music clip. Usually k-means clustering or GMM is used to fit the underlying distribution of the features. The music similarity of two songs is calculated as the statistical distance between the feature distributions of the two songs. As noted in Section 1, the previous methods mentioned above have several shortcomings associated to the iterative fitting of a mixture model and the computation of the pairwise distance. In contrast, the closed-form vector representation of the Fisher kernel in [5] can be readily extended to represent the auditory features and easily incorporated with simple distance measures, such as Euclidean or Cosine distance. We introduce the Fisher kernel in Section 2.1 and apply it to the music similarity in Section 2.2.

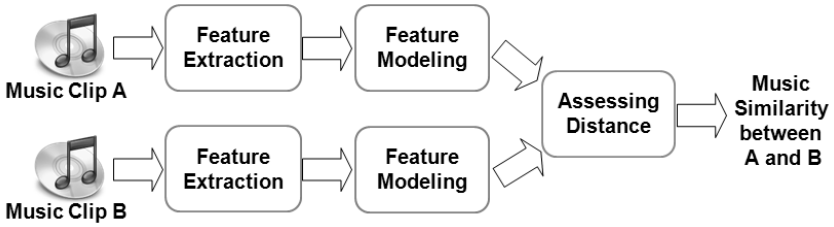


Fig. 1. Overview of the content-based music similarity computation.

2.1 Fisher Kernel

The followings are the introduction to the Fisher kernel as was proposed in [4][5]. Let X be a sample whose generation process can be modeled by a probability density function p with parameters λ [6]. In this paper, X corresponds to feature vectors from a music clip. With respect to the parameters λ, the gradient vector of X is denoted by

$$G_\lambda^X = \nabla_\lambda \log p(X|\lambda) \ . \tag{1}$$

The gradient vector gives the direction in parameter space into which the learnt distribution should be modified to better fit the observed data [8]. The dimensionality of this vector depends only on the number of parameters in λ [5]. On the gradient vector, a kernel is defined in [4][5][6] in the inner product form as follows:

$$K(X,Y) = G_\lambda^X F_\lambda^{-1} G_\lambda^Y \tag{2}$$

where F_λ is the Fisher information matrix of p given by

$$F_\lambda = E_{x\sim p}\left[\nabla_\lambda \log p(x|\lambda) \nabla_\lambda \log p(x|\lambda)'\right] \ . \tag{3}$$

Through the Cholesky decomposition of $K(X,Y)$, a normalized gradient vector [5] is obtained as follows:

$$\varrho_\lambda^X = F_\lambda^{-1/2} \nabla_\lambda \log p(X|\lambda) \tag{4}$$

The normalized gradient is referred to as the Fisher vector of X [5] which is the gradient of the sample's likelihood with respect to the parameters of the underlying distribution, scaled by the inverse square root of the Fisher information matrix [8].

2.2 Music-Similarity Computation Based on the Fisher vector

As shown in Fig. 2, we first extract the low-level spectral features from an input audio. An audio signal is split into overlapping segments (called frames) of length L with 50% overlap (in our system, $L = 1024$ at a sampling frequency of 22050

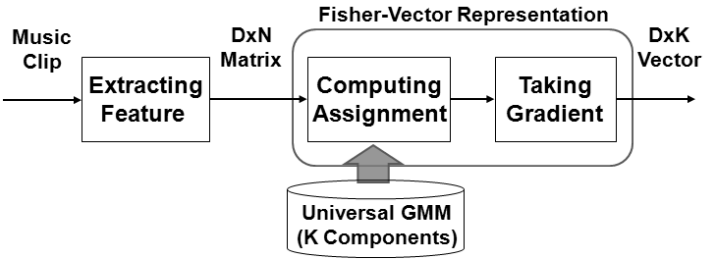


Fig. 2. Extraction of the Fisher vector from a music clip.

Hz). Each frame is windowed by a Hamming window of length L and transformed into the frequency domain. From each frame, we extract the low-level spectral features. We consider the D-order MFCC (in this paper, $D = 19$) as the low-level spectral feature as in [1]. Assuming that there are N frames in a music clip, the set of MFCC vectors from each frame is given by $X = \{x_0, x_1, \cdots, x_{N-1}\}$. We choose the GMM as a underlying distribution p for the feature space since the GMM has been used to represent the MFCC space in [2][3]. We denote the distribution p as a sum of mixtures by $p(x) = \sum_{k=0}^{K-1} w_k N(x|m_k, \Sigma_k)$ where the mixture weight w_k, mean vector m_k, and covariance matrix Σ_k are the parameters λ. In order to simplify the representation, as in [5], the covariance matrix is constrained to be diagonal with variance vector σ_k^2. We only consider the Fisher vector with respect to the mean and the standard deviation since that with respect to the weight carries little information [6]. Based on the assumption that x_n's are generated independently from p [6], the Fisher vector with respect to a parameter λ is given by

$$G_\lambda^X = \frac{1}{N} \sum_{n=0}^{N-1} \nabla_\lambda \log p(x_n|\lambda) \quad . \tag{5}$$

A closed-form expression of the Fisher information matrix of a GMM was derived in [5]. Using the derived Fisher information matrix, the Fisher vector ϱ_μ^X for the mean and ϱ_σ^X for the standard deviation are simplified in [5][6][7] as follows:

$$\varrho_\mu^X[kD + d] = \frac{1}{N\sqrt{w_k}} \sum_{n=0}^{N-1} \gamma_{nk} \left(\frac{x_n[d] - \mu_n[d]}{\sigma_n[d]} \right) \tag{6}$$

$$\varrho_\sigma^X[kD + d] = \frac{1}{N\sqrt{2w_k}} \sum_{n=0}^{N-1} \gamma_{nk} \left[\left(\frac{x_n[d] - \mu_n[d]}{\sigma_n[d]} \right)^2 - 1 \right] \tag{7}$$

where d denotes the d-th dimension of the feature vector x_n (in our case, $d = 0, 1, 2, \cdots, D-1$, and $k = 0, 1, 2, \cdots, K-1$), and γ_{nk} is the soft alignment (posterior probability) of feature vector x_n to the k-th Gaussian component of

the GMM given by

$$\gamma_{nk} = \frac{w_k N(x_n|m_k, \sigma_k)}{\sum_{j=0}^{K-1} N(x_n|m_j, \sigma_j)} \quad . \tag{8}$$

As shown in the Fig. 2, the Fisher kernel transforms an incoming variable-size (in our case, $D \times N$) set of independent features into a fixed-size (in our case, $D \times K$) vector representation, assuming that the features follow a parametric generative model estimated on a training set [8].

3 Experimental Results

Evaluating a music similarity function is intricate since the ground truth of the music similarity is difficult to obtain. Thus, in the previous works [1][2], it was assumed that the songs of the same genre or singer are perceptually more similar than those of the different genre or singer. With the same assumption, we evaluate the validity of the Fisher vector for music similarity on the genre and the singer datasets. The genre dataset is made by George Tzanetakis for his work [9] and consists of 1000 songs over ten different genres: blues, classical, country, disco, hiphop, jazz, metal, pop, reggae, and rock. The singer dataset is made by the authors and consists of 680 songs (20 songs per each singer) over 34 singers. For each query song in the dataset, we calculate the distances with the other songs in the dataset and examine the closest 5, 10, and 20 songs among which we count the number of songs in the same category (genre or singer) as the query song. The Fisher-vector based music similarity is compared to the Logan's music similarity function [1], where the MFCC vectors extracted from a song are modeled by the k-means clusters, and the clusters from two songs are compared each other using the EMD [1].

Each song in the datasets was converted to mono at a sampling frequency of 22050 Hz and then divided into frames of 46.4 ms ($L = 1024$) overlapped by 23.2 ms. We computed the 19-order MFCC of each frame as a low-level feature ($D = 19$). When extracting the Fisher vector, we considered three different GMMs as the underlying feature distribution with the number of mixture components in the GMM as 4, 8, and 16. The GMM was trained on 156 songs of various genres which are not overlapping with the test datasets. For each song, we calculated the Fisher vector with respect to the mean and the standard deviation as in (6) and (7) respectively. Table 1 is the result of the genre dataset and shows the average number of closest songs with the same genre as the query song. Table 2 is the result of the singer dataset and shows the average number of closest songs by the same singer as the query song. In obtaining the results in Table 1 and 2, each song in the dataset was used as a query, and the closest 5, 10, and 20 songs to each query were scrutinized. On the genre dataset (10 genres), the expected number of songs with the same genre as the query song among the closest 5 songs is $0.5 (= 5 \times 1/10)$ for random selection (assuming the identical and independent trials). In case of the singer dataset (34 singers),

Table 1. Average number of closest songs with the same genre as the seed song. The MFV and the SFV denote the Fisher vector with respect to the mean and the standard deviation respectively.

Types of Signatures	Distance Measure	Average number of songs in the same genre		
		Closest 5	Closest 10	Closest 20
MFV ϱ_{μ}^{X} $(K = 4)$	Euclidean	2.492	4.325	7.501
	Cosine	2.052	3.754	6.740
SFV ϱ_{σ}^{X} $(K = 4)$	Euclidean	1.054	1.979	3.73
	Cosine	1.569	2.846	5.16
MFV ϱ_{μ}^{X} $(K = 8)$	Euclidean	2.483	4.329	7.368
	Cosine	2.314	4.106	7.436
SFV ϱ_{σ}^{X} $(K = 8)$	Euclidean	1.463	2.728	5.012
	Cosine	1.643	2.965	5.391
MFV ϱ_{μ}^{X} $(K = 16)$	Euclidean	2.482	4.352	7.484
	Cosine	**2.545**	**4.557**	**8.014**
SFV ϱ_{σ}^{X} $(K = 16)$	Euclidean	1.581	2.929	5.358
	Cosine	1.640	2.994	5.487
Logan's Method [1]	EMD	**2.743**	**4.801**	**8.384**
Random Selection		0.5	1.0	2.0

Table 2. Average number of closest songs by the same singer as the seed song. The MFV and the SFV denote the Fisher vector with respect to the mean and the standard deviation respectively.

Types of Signatures	Distance Measure	Average number of songs by the same singer		
		Closest 5	Closest 10	Closest 20
MFV ϱ_{μ}^{X} $(K = 4)$	Euclidean	1.663	2.749	4.118
	Cosine	0.726	1.229	2.116
SFV ϱ_{σ}^{X} $(K = 4)$	Euclidean	0.319	0.602	1.096
	Cosine	0.559	0.929	1.497
MFV ϱ_{μ}^{X} $(K = 8)$	Euclidean	1.713	2.756	4.126
	Cosine	1.326	2.290	3.631
SFV ϱ_{σ}^{X} $(K = 8)$	Euclidean	0.476	0.804	1.410
	Cosine	0.547	0.929	1.531
MFV ϱ_{μ}^{X} $(K = 16)$	Euclidean	1.790	2.912	4.313
	Cosine	**1.919**	**3.121**	**4.800**
SFV ϱ_{σ}^{X} $(K = 16)$	Euclidean	0.618	1.091	1.890
	Cosine	0.656	1.151	1.999
Logan's Method [1]	EMD	**1.743**	**2.776**	**4.044**
Random Selection		0.147	0.294	0.588

the expected number of songs by the same singer as the query song among the closest 5 songs is $0.147(= 5 \times 1/34)$ for random selection. These indicate that the feature-based music similarity could provide a playlist which is much more meaningful than the random shuffling. In both Table 1 and 2, the Fisher vector with respect to the mean outperformed that with respect to the standard deviation. As the number of GMM components got larger (i.e. the dimensionality of the Fisher vector increased), the retrieval performance improved gradually. However, the performance gain was not quite notable. The retrieval performance of the Fisher-vector representation was more or less similar to that of the Logan's method [1] for both datasets. We note that the Fisher-vector representation has several merits over the Logan's method as stated in the Section 1. Moreover, the Fisher-vector representation is in vector form where many kinds of the distance measures can be easily incorporated. Although the Euclidean and the Cosine distance are considered in this paper, other distance measures can also be employed for the Fisher vector to boost the retrieval performance further. We leave it as a future work. We note that the scope of the experimental results in this paper is limited to the objective relevance with respect to the genre and the singer criterion. Each person's basis of the music similarity is multifarious depending on the personal preference and familiarity to a certain type of music [10]. Since designing and performing a subjective test on the music similarity is quite intricate in practice [10][11], we focus on the comparison between the proposed and the Logan's approach [1] with two objective criterions: the genre and the singer metadata. Further investigations of the proposed music similarity function are necessary with a subjective criterion by the empirical ratings of the human listeners to complement the experimental results reported in this paper.

4 Summary

In this paper, we apply the Fisher-vector representation of the spectral features to the content-based music similarity computation. The distance between the Fisher vectors of two songs is used as the similarity of the two songs. Compared with the previous mixture model representation, the Fisher-vector representation could provide a simplified alternative framework for music similarity computation. Experimental results show that the Fisher-vector representation can match the retrieval performance of the more complex ones.

References

1. Logan, B., Salomon, A.: A Music Similarity Function Based on Signal Analysis. In: IEEE International Conference on Multimedia and Expo, pp. 745–748. Tokyo (2001)
2. Aucouturier, J.-J., Pachet, F.: Improving Timbre Similarity : How High's the Sky?. Journal of Negative Results in Speech and Audio Sciences. 1, (2004)
3. Mandel, M., Ellis, D.: Song-Level Features and Support Vector Machines for Music Classification. In: International Conference on Music Info. Retrieval, London (2005)

4. Jaakkola, T., Haussler, D.: Exploiting generative models in discriminative classiers. In: Advances in Neural Information Processing Systems, pp. 487–493. Vancouver (1999)
5. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: IEEE Computer Vision and Pattern Recognition, pp. 1–8. Minneapolis (2007)
6. Perronnin, F., Sanchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: European Conference on Computer Vision, pp. 143–156. Crete (2010)
7. Perronnin, F., Liu, Y., Sanchez, J., Poirier, H.: Large-scale image retrieval with compressed Fisher vectors. In: IEEE Computer Vision and Pattern Recognition, pp. 3384–3391. San Francisco (2010)
8. Jegou, H., Douze, M., Schmid, C., Perez, P.: Aggregating local descriptors into a compact image representation. In: IEEE Computer Vision and Pattern Recognition, pp. 3304–3311. San Francisco (2010)
9. Tzanetakis, G., Cook, P.: Musical genre classification of audio signals. IEEE Transaction on Speech and Audio Processing. 10, 293–302 (2002)
10. Lee, J.H.: How similar is too similar?: Exploring users' perceptions of similarity in playlist evaluation. In: International Conference on Music Info. Retrieval, Miami (2011)
11. Bogdanov, D., Herrera, P.: How much metadata do we need in music recommendation? A subjective evaluation using preference sets. In: International Conference on Music Info. Retrieval, Miami (2011)

Automatic Performance of *Black and White n.2*: The Influence of Emotions Over Aleatoric Music

Stefano Baldan, Adriano Baratè, and Luca A. Ludovico

LIM - Laboratorio di Informatica Musicale
Dipartimento di Informatica
Università degli Studi di Milano
Via Comelico 39/41, I-20135 Milano, Italy
{baldan,barate,ludovico}@dico.unimi.it

Abstract. *Black and White n.2* is a piece of aleatoric music by Franco Donatoni. Conceived as a set of 120 exercises for piano, it uses a non-conventional way to encode the score. Some elements of the composition are left to chance, thus they should be extemporarily determined by the performer. In this work, human choices are performed by an automatic system. The algorithms designed and described here extract emotion-related information from a video input and consequently create in real time an instance of the piece. Finally, the paper presents the case study of *Black and Byte*, an application implemented to test such algorithms.

Keywords: aleatoric music, automatic composition, emotions

1 Introduction

The relationships between music and emotions is a very rich and complex matter, and a theoretical discussion of such a subject goes beyond the goals of the present paper. In the computer field, this problem has been addressed for instance in [1], [2] and [3].

This work narrows the field by focusing on the relationships between music performance and emotions. Even in a traditional context, such as an evening at the concert hall, a music performance is influenced by the feelings of the performers and it conveys emotions to the audience. From this point of view, in many contemporary music pieces and multimedia installations new frontiers have been explored, making the listener become the protagonist of the performance. For instance, the emotions and behaviours of the audience during a performance can be captured in order to influence the performance itself in real time.

Our work is strictly related to the latter aspect. The goal is automatically rendering an aleatory composition for keyboard (see Section 2) through a computer-based system able to solve its non-determinism. In order to generate an automatic performance, all the values of aleatory variables must be computed. In general it could be sufficient to generate sequences of pseudo-random numbers, if necessary under certain conditions. But for our purposes at least two further constraints must be considered:

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

1. The aleatory aspects of the performance should not be fully determined by chance, but influenced in real time by emotion-related contents. In particular, our algorithms take both a score and a video as input. Some features of the video are automatically extracted and evaluated in order to add a coherent soundtrack based on the score;
2. The resulting performance for keyboard instruments, after a transcription in conventional notation, should be playable by a medium-skilled human performer. This aspect is not trivial, since generated chords have to respect the fingering rules explicitly indicated in the score, and to consider hand posture and comfort.

All these items will be discussed in detail in the following sections.

2 F. Donatoni's *Black and White n.2*

Black and White n.2 [4] is a collection of 120 pieces written by the Italian composer Franco Donatoni (9 June 1927 - 17 August 2000). They can be played on any keyboard instrument, including piano, harpsichord, celesta, electronic keyboard, etc. Versions for 2 and 3 keyboard instruments have been conceived and performed as well. The subtitle of the composition is *Esercizi per le 10 dita*, namely 10-finger exercises, and this aspect will be fundamental to understand Donatoni's notation, as explained below. This piece belongs to the genre known as *aleatoric music* [5], since some primary parameters of the composition are not predetermined, but their values depend on random processes or extemporary decisions made by the performer.

In the preface to the score, the author briefly explains the simple set of rules to read the score, whose conventions significantly differ from Common Western Notation (CWN). In fact, the two staves usually assigned to traditional keyboard notation (i.e. the grand staff) in this case do not carry pitch and rhythm, but rather finger-related information. Only lines are used, and each line corresponds to a specific finger. For the right hand, the lower line corresponds to the thumb and the upper line to the little finger, and vice versa for the left hand. Consequently, the typical symbols of a traditional score are not present: no time nor key indication, no bars nor rhytmic values, etc.

As mentioned before, a key constraint is the use of one or more keyboard instruments. Given this hypothesis, the rules to read the score are few:

- The association between symbol positions over lines and fingers is fixed;
- About the colour of the symbols, each symbol can be either filled or empty. This graphical convention forces the performer to play either a black or a white key respectively;
- The symbols that can be placed over this sort of staff have either a circle or a square shape, and this aspect is related to dynamics. At the beginning of the performance the player decides if circles should correspond to the *ppp* dynamic indication (i.e. softest possible), thus squares would correspond to *fff* (i.e. loudest possible), or vice versa;

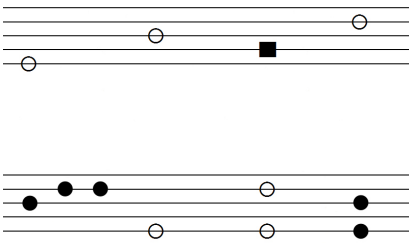


Fig. 1. A score excerpt which follows Donatoni's *Black and White n.2* set of rules.

- The concept of chord is associated to the vertical alignment of such symbols, possibly spanning over the two staves;
- Finally, arrows pointing up or down can be specified for each chord, even a degenerate 1-note chord. The meaning of either an upward or a downward arrow is using preferably either the higher or the lower octaves of the keyboard respectively.

The application of this set of rules clearly leaves many music parameters to the determination of the performer. As a consequence, for a given score a number of different performances is possible. In particular, the following parameters can be considered as degrees of freedom:

- Metronome tempo and time indication are not present. Consequently, the piece is not organized into bars, nor into other regular rhythmical grouping;
- All rhythmic aspects (note durations, note density, articulation, etc.) are not specified. As regards articulation, Donatoni allows the use of *legato*, *staccato* and *tenuto*, as well as a free use of sustain pedal;
- Specific pitches are not indicated. Score information about pitches forces only the use of either white or black keys in a given number, and provides suggestions about the pitch range to use (i.e. the lower, middle or higher octaves over the keyboard).

3 Extraction of Emotional Features from Video

As Donatoni explicitly states in the preface to *Black and White n.2*, this composition is only partially defined by a limited set of rules, while many other aspects are left to the extemporaneous interpretation and execution by the single performer. Such a work can therefore be heavily influenced by the mood of the musicians who are playing it, and on the other side it can also be heavily modified as regards melody, rhythm and harmony in order to transmit a certain kind of emotion.

The great affective versatility of *Black and White n.2* could be exploited to automatically generate consistent soundtracks for arbitrary video footage. In this

paper we will propose simple yet effective methods to extract affective features from motion videos, both in online and offline scenarios. Then a possible mapping of the rules of Donatoni's composition will be provided in order to synchronize an automatic musical performance with the visual information in terms of timing and emotions.

3.1 The Affective Model: Related Work

Before extracting the emotional features out of the video footage, we must find a way to describe, analyze and measure those ephemeral entities we call "emotions". To face this non-trivial problem, an interdisciplinary research in heterogeneous fields is required, including Music Psychology and Music Information Retrieval. A review of some of the most relevant contributions makes two different approaches emerge: *categorical analysis* versus *dimensional analysis*.

In the former case, emotions are defined in a discrete way and classified as belonging to a few basic categories, such as "love", "hate", "joy", "sorrow" and so on (see [6] for further details). As stated in [7], one of the difficulties in the automatic recognition of emotions is labeling of the data. Many experiments have been conducted in this sense, starting either from symbolic contents or from audio objects. It is worth citing the data labeling proposed by Hevner in 1936, which consists of a circle of 8 classes where not all adjectives in a single group are synonims [8]. A more recent work in this context is [9], which starts from the previous one and proposes 13 classes each labeled by one, two or three adjectives. It should be clear that adding classes and dimensions means moving from a discrete to a continuous description. For our purposes, one of the main drawbacks of categorical analysis lies in its discrete nature, which does not catch the subtle nuances of human feelings in a satisfactory way. Moreover, this model describes emotions qualitatively, making it difficult to map them onto the quantitative parameters used to generate an automatic musical performance, like note pitches, beats per minute and so on.

In dimensional analysis, on the contrary, emotions are defined inside a multidimensional, continuous space. An interesting work that applies this kind of analysis to emotions detection in speech is [10], which adopts the activation dimension. In the field of affective retrieval of information, the Arousal-Valence (AV) model proposed in [11] is still considered up-to-date and it is one of the most used. For example, in [12] it is applied to problems of music classification by affective contents. The model defines a two-dimensional space where emotions are classified in terms of their level of arousal (calm versus excited) and pleasantness (positive versus negative). The dimensional approach gives the possibility to express the subtle nuances of human feelings through continuous variations in the chosen multidimensional space. Since it is based on quantitative features, it also allows an easy mapping onto the musical parameters used to generate the automatic performance. For these reasons, the present work applies a dimensional analysis technique, and in particular the AV model.

3.2 Arousal and Valence in Videos

After choosing an affective model, the next step is identifying the emotional features in motion video information, namely which parameters are relevant to the arousal and valence perception of what we see. In a previous work [13], Zhang proposes an affective categorization of musical videoclips based on five key features: sequence changes, motion, lighting, color saturation and color energy. The first two features are related to the arousal dimension, while the remaining ones are related to the valence dimension.

The work shows how a high number of sequence changes and a high amount of moving objects, camera zooming and panning effects result in a sense of excitation, whereas few scene cuts and static planes give a sense of quietness. On the other side, vivid colours and bright images are common in calm or joyful videos, while faded and dark images are often used to convey a sense of sadness, fear or anger.

In the proposed affective analysis of musical videoclips, also audio clues are taken into account. From this point of view, the key features are zero-crossing rate, tempo and beat strenght for a classification along the arousal dimension, and rhythm regularity and pitch as regards the valence dimension. While the work by Zhang focuses on the extraction of these parameters out of the existent soundtracks of musical videoclips, our goal is exactly the opposite, namely to synthesize a soundtrack whose audio clues are consistent with the ones extracted by the visual information. There are some significant differences in our approach: while the work of Zhang focuses on the offline affective analysis of a database of MPEG musical videoclips, we want to generate soundtracks for arbitrary video footage, stored in different encodings or even acquired in real time by a webcam or a RTP[1] stream.

3.3 Algorithms for Video Analysis

The algorithms used to detect sequence changes takes inspiration from the abrupt scene change detection described in [14]. It is a method based on inter-frame motion intensities, which can also be directly employed to detect motion in videos. The metric used to compute inter-frame motion intensities is the absolute difference between consecutive frames, which is described by the formula:

$$D = \sum_{x=0}^{W} \sum_{y=0}^{H} |r_1(x,y) - r_0(x,y)| + |g_1(x,y) - g_0(x,y)| + |b_1(x,y) - b_0(x,y)|$$

where W and H are respectively the width and height of both frames, $r_1(x,y)$, $g_1(x,y)$ and $b_1(x,y)$ are the red, green and blue values of the pixel at coordinates (x,y) inside the current frame, and $r_0(x,y)$, $g_0(x,y)$ and $b_0(x,y)$ are their corresponding values inside the previous frame. Scene-change detection is achieved by keeping trace of the maximum motion intensity occurring inside a local window of consecutive frames, and comparing the motion intensity of every frame

[1] RTP stands for Real Time Protocol, which is described in RFC 3550.

against this value. If the result for the current frame is n times the computed maximum, then a sequence change is detected. According to Yeo, values of n between 2 and 3 have proven to give good results. The local moving frame-window prevents the detection of false positives such as camera flashes or rapid panning and zooming of the scene. A window of 25 frames for a 25 fps video, for example, means that there cannot be two consecutive sequence changes within a second. Once sequence changes are detected, the number of shots per second can be determined dividing the number of shots in the video by its duration in seconds. In the present work this value is computed locally instead for the whole video, choosing a window of 5 sequence shots, as we want to capture in real time the local variations of this feature inside each video stream.

Valence features can be easily obtained by converting pixel values from the RGB into the HSB color space.[2] Brightness is computed in a straightforward way summing the brightness of pixels in each single frame. Similarly, saturation is computed summing the saturation of pixels in each single frame. Color energy instead is a composite parameter obtained from the combination of the two former values with the standard deviation of the pixel hue values. The underlying concept is that colorful videos often contain many different tonalities, thus yielding very high standard deviation values for the hue; on the contrary, faded videos are more likely to contain a limited amount of different colors, thus yielding low standard deviation values for this parameter.

All the key features are normalized for convenience between 0 and 1: the value of shots per second is already in this range if a window of one second (usually 25 or 30 frames, depending on the standards and formats adopted) for scene change detection is chosen. On the other side, inter-frame motion is divided by the maximum possible absolute difference between frames, which for a 24 bit RGB color video is:

$$\hat{D} = \frac{D}{W * H * 255 * 3}$$

Similarly, lighting and saturation are divided by their maximum possible values inside a single frame, which for a 24 bit HSB color video is:

$$\hat{L} = \frac{L}{W * H * 255}$$

$$\hat{S} = \frac{S}{W * H * 255}$$

Finally, the standard deviation of the hue is normalized using the formula:

$$\hat{\sigma}_h^2 = \begin{cases} 4\dfrac{\sigma_h^2}{h_{max}}, & \text{if } \sigma_h^2 < \dfrac{h_{max}}{4} \\ 2 - 4\dfrac{\sigma_H^2}{h_{max}}, & \text{otherwise} \end{cases}$$

[2] The RGB acronym represents an additive color model in which red, green, and blue light are added together to reproduce colors. HSB is another color model, based on the image attributes called hue, saturation, and brightness.

The final arousal value is obtained by a user-defined weighted average between shots per second and motion intensity. Similarly, the valence feature is computed by a user-defined weighted average between lighting, saturation and hue standard deviation. Both arousal and valence can be then amplified and clipped to the maximum value of 1. These manual adaptations are included in order to make the system adaptable to changes in lighting, input devices, nature of the video footage and so on.

4 Mapping Emotional Features onto a Score

The final goal of our work is the extraction of emotions from video captures in order to drive a computer-based performance of *Black and White n.2*. As discussed in Section 3, the algorithm to extract emotion values from a video employs a 2-axes classification, based on arousal and valence values respectively. Now we will explain which music features can be used, and how, in order to convey emotions and consequently to create an adequate soundtrack for motion videos. Inspiring works from this point of view are [15] and [16]. Please note that, in order to establish adequate mappings between music-conveyed emotions and musical features, psychology and neuropsychology studies must be considered, too (e.g. [17] and [18]).

Some choices specific for our implementation will be detailed in Section 5.

4.1 Rhythm

Starting from mentioned research, rhythmic aspects have been mapped onto the arousal dimension.

Note durations are not expressed as in traditional music theory (e.g. crotchets, quavers, etc.), but as absolute time intervals whose value belongs to a continuous range. No time indication or BPM[3] value is provided by the score, and no one is introduced by our algorithms.

Other aspects related to rhythm are articulation marks and pedals. The preface to Black and White n.2 explicitly cites the possibility to introduce articulations and pedals in a performance, even if the corresponding marks do not belong to allowed score symbols. As regards in-use articulation signs, they usually include *staccatissimo*, *staccato*, *martellato*, *marcato*, *tenuto*, and so on. Besides, a *legato* effect is usually indicated through slurs. Finally, the damper pedal (if present on the keyboard instrument) represents the maximum level of sustain, as all notes played will continue to sound until the pedal is released. All these aspects, which are allowed by Donatoni but not encoded inside the score, are left to improvisation and to the performer's feelings. In our opinion, the list provided before should correspond to a progressive reduction of the arousal level, ranging from *staccatissimo* to *legato* with sustain pedal.

[3] BPM stands for Beats Per Minute.

4.2 Harmony

Harmony, here intended as a sequence of chords, is one of the most important, but also one of the most difficult music dimensions to map onto the mentioned axes. Problems arise for a number of different reasons.

First, the function of a chord changes depending on the surrounding context, so that an evaluation should involve not only the chord itself, but also the harmonic path it has been inserted into. For example, a major triad is usually considered happier and brighter than a minor one, but in a tonal context a minor triad built on the second degree of a major scale (e.g. [D,F,A] in C major) does not necessarily convey a sense of sadness.

Other chords are intentionally ambiguous and their meaning is clarified only by the surrounding harmonic path. A very significant example is the *incipit* of the *Allegro non molto* from Violin Concerto in F minor RV 297 "L'inverno" by A. Vivaldi. The first chord (see Figure 2) is incrementally built by superimposing higher and higher voices: initially it sounds affirmative and stable (the F-minor tonic alone), then ambiguous (major second), misleading (minor sixth), frightful (perfect fourth), and finally clear thanks to the resolution of the dissonance on a diminished seventh chord.

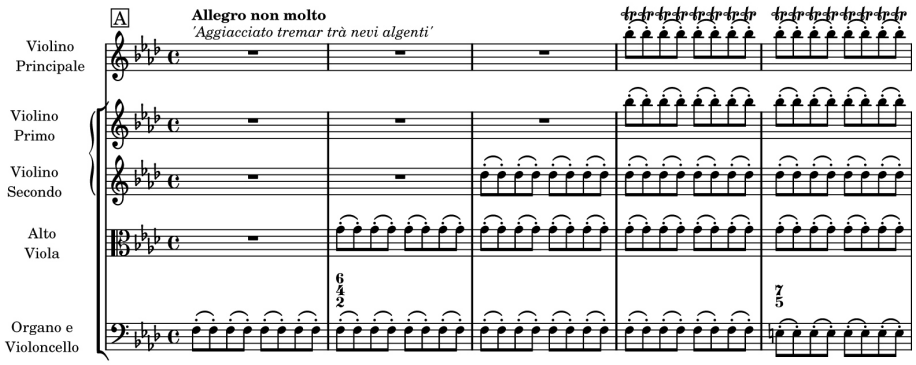


Fig. 2. *Incipit* of the *Allegro non molto* from Violin Concerto in F minor RV 297 "L'inverno" by A. Vivaldi.

Even those chords that present a commonly-accepted affective value in Western culture do not have an implicit nor a universal value, above all in a non-tonal context. An interesting survey about harmony and chord functions in the twentieth-century music can be found in [19].

Obviously, an exhaustive discussion of the matter goes beyond the goals of the present paper. Now we will briefly explain how the problem has been solved in our work from a practical perspective. Let us recall the two key concerns:

1. It is necessary to find an efficient and effective way to map in real time emotional features onto AV axes. In our work, the harmonic path is not predetermined: the performer has to chose pitches extemporarily, as the author himself recommends. A reductive but practical solution is considering chords as isolated entities, thus ignoring the harmonic context;
2. Fingering indications are one of the few contraints provided by the score, and they cannot be ignored. As a consequence, not all chords that prove to be adequate as regards their affective characteristics can be employed, but only those which support a given fingering.

Our approach can be described as a four-steps process.

First, the system is provided with a set of chord models to use, including their possible inversions. The concept of "model" means encoding halftone distances from the root note, instead of any possible combination of pitches. An easy-to-implement algorithm can produce any instance of a model starting from each available pitch.

Then, each chord model is put in correspondence with two ranges of continuous values, on the arousal and on the valence axis respectively, thus forming a rectangle. In this way, any valid point of the AV plane is covered by a variable number of overlapping rectangles, corresponding to all the chord models that can convey those given arousal and valence senses.

After defining the set of suitable chord models for a given pair of AV values, a further selection is made on the base of fingerings. In fact, not all the chords may support a particular hand position.

Finally, it is necessary to verify if the selected chord model has at least one instance that corresponds to the black/white keys configuration indicated by the score.

If one of the mentioned steps fails, as there exist no candidates having the required chracteristics, a backtracking technique is used to select a new candidate. For example, if a major triad, a dominant seventh and a diminished seventh cover the current point of the AV plane and the score requires a 4-fingers chord, let the second chord be initially selected as the candidate. After verifying that the required black/white layout cannot be instanced starting from the dominant seventh model, the algorithm selects the diminished seventh.

Please note that a formal check must be conducted on the chord-models set to verify the complete covering of the AV plane using all possible fingerings and black/white-key configurations.

As regards our implementation, details about chord models and single-chord mappings onto the AV plane will be provided in Section 5.

5 Case study: *Black and Byte*

Black and Byte is the application designed and implemented to test the algorithms described in Sections 3 and 4.

The interface allows to open a score in plain-text format. The file provides a score view on a single staff system, which is usually defined "scroll view" in

music editing software. The two "staves" corresponding to right and left hand are separated by an empty line. Inside such a document, the allowed symbols are: empty circle ○, filled circle ●, empty rectangle □, filled rectangle ■, up arrow ↑, and down arrow ↓. Whitespaces and tabs are supported as well, but they have no musical meaning. Other symbols are not managed, so they are ignored by the parser.

As regards video input, it is possible either to load an available media file or to acquire motion images from a webcam or a stream in real time. The audio track (if available) is ignored. Since the duration of the performance is not known in advance (this is one of the aspects left to the performer's will), score is read in a circular way and a loop is performed to sonorize the entire video. At each iteration, music parameters are recalculated according to Donatoni's rules.

Some features of the prototype are not hard-coded, so they can be configured by the user. It is worth citing the chord-model list, which deeply influences the results of our algorithms. At present, the total amount of supported chord models is limited to those typical of traditional harmony, namely:

1. 10 bichords (minor second, major second, minor third, major third, diminished fourth, perfect fourth, augmented fourth, diminished octave, perfect octave, augmented octave) and their 10 inversions;
2. 7 triads (major, minor, diminished, augmented, suspended second, suspended fourth, and flat fifth) and their 14 inversions;
3. 7 sevenths (major, minor, dominant, diminished, half-diminished, minor/major, and augmented/major) and their 21 inversions;
4. 7 ninths (ninth, minor ninth, flat ninth, minor flat ninth, augmented ninth, nine-six, minor nine-six) and their 28 inversions.

Inverted chords are automatically computed from the model of the "parent" chord, namely the root-position chord. Since chord-model list has not been hard-coded inside the prototype, this set could be easily extended, for instance supporting elevenths and thirteenths, post-tonal chords or micro-tonal intervals.

The mapping of chords onto the AV plane is clearly subjective. In our approach:

- Consonance/dissonance among chord notes correspond to a low/high level of arousal, respectively. This parameter is computed by evaluating the intervals inside the chord, namely the relationship between the root element and the following notes;
- The belonging of the chord to the minor/major tonal area corresponds to a low/high level of valence.

For instance, a major triad is very consonant and clearly belongs to the major area, so it has a low value for arousal and a high value for valence: it can convey a sense of brightness, grace, quietness, solemnity, etc. On the other side, ambiguous chords - e.g. empty fifths - present a neutral value on both dimensions: they can represent bore, doubt, indefiniteness, etc.

The interface presents two key elements: a media player where the motion video is loaded, and a panel containing the original score. As mentioned before, the creation of the performance is extemporary, and it is computed in real time depending on video analysis. While the performance is advancing, the interface shows the corresponding transcription in CWN together with fingering indications. In this way, on one side the instances of Donatoni's rules can be verified, and on the other side the resulting score can be played also by a human performer.

A component not directly related to the real-time performance, but useful to show video-analysis results and to understand the consequent sonorization, is the panel containing the AV plane representation. A small circle defines the position of the current point along the arousal and valence axes.

6 Conclusions

In this paper we have proposed a process to sonorize motion videos in real time, starting from 1) an on-the-fly analysis of video contents and 2) a given score encoded according to Donatoni's rules.

One of the goals was exquisitely theoretical: testing dimensional-analysis results and determining efficient and effective algorithms to apply those results to aleatoric music. Besides, a number of practical applications can exist, e.g. in multimedia installations, emotion-based sonorization of videos, and real-time control of aleatory performances through face recognition or other gestures.

Acknowledgments. This work has been partially funded by the Enhanced Music Interactive Platform for Internet User (EMIPIU) project.

References

1. Yang, D., Lee, W.S.: Disambiguating music emotion using software agents. In: Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR04), pp. 52–58 (2004)
2. Yang, Y.H., Lin, Y.C., Su, Y.F., Chen, H.H.: A regression approach to music emotion recognition. IEEE Transactions on Audio, Speech, and Language Processing, vol. 16(2), pp. 448–457. IEEE Press, New York (2008)
3. Livingstone, S.R., Brown, A.R.: Dynamic response: real-time adaptation for music emotion. In: second Australasian conference on Interactive entertainment, Proceedings of the, pp. 105–111. Creativity & Cognition Studios Press (2005)
4. Donatoni, F.: Black and white II - Esercizi per le dieci dita per strumenti a tastiera. Suvini Zerboni, Milano (1968)
5. Meyer-Eppler, W.: Statistic and Psychologic Problems of Sound. Die Reihe No. 1: Electronic Music, pp. 55–61 (1958)
6. Plutchik, R.: The Nature of Emotions. American Scientist, vol. 89(4), pp. 344–350 (2001)
7. Wieczorkowska, A., Synak, P., Lewis, R., W. Raś, Z.: Extracting emotions from music data. Foundations of Intelligent Systems, pp. 456–465. Springer (2005)

8. Hevner, K.: Experimental studies of the elements of expression in music. American Journal of Psychology, n. 48, pp. 246-268 (1936)
9. Li, T., Ogihara, M.: Detecting emotion in music. In: 4th International Conference on Music Information Retrieval ISMIR, Proceedings of the, pp. 239–240 (2003)
10. Tato, R., Santos, R., Kompe, R., Pardo, J.M., Emotional space improves emotion recognition. Seventh International Conference on Spoken Language Processing (2002)
11. Schlosberg, H.: Three dimensions of emotion. Psychological review, vol. 61(2), pp. 81–88. American Psychological Association (1954)
12. Oliveira, A., Cardoso, A.: Towards bidimensional classification of symbolic music by affective content. In: International Computer Music Conference, Proceedings of the (2008)
13. Zhang, S., Huang, Q., Jiang, S., Gao, W., Tian, Q.: Affective visualization and retrieval for music video. IEEE Transactions on Multimedia, vol. 12(6), pp 510–522. IEEE Press, New York (2010)
14. Yeo, B.L., Liu, B.: A unified approach to temporal segmentation of motion JPEG and MPEG compressed video. In: Multimedia Computing and Systems, Proceedings of the International Conference on, pp. 81–88. IEEE Press, New York (1995)
15. Gabrielsson, A., Lindström, E.: The influence of musical structure on emotional expression. Oxford University Press (2001)
16. Wu, T., Jeng, S.: Extraction of segments of significant emotional expressions in music. In: 2006 International Workshop on Computer Music and Audio Technology, Proceedings of the, pp. 76–80 (2006)
17. Marin, O.S.M., Perry, D.W.: Neurological aspects of music perception and performance. Academic Press (1999)
18. Simpson, J.A.: Music and the Brain - Studies in the Neurology of Music. Journal of Neurology, Neurosurgery & Psychiatry, vol. 40(7). BMJ Publishing Group Ltd. (1977)
19. Persichetti, V.: Twentieth-century harmony: creative aspects and practice. WW Norton (1961)

The Visual SDIF interface in PWGL

Mika Kuuskankare*

Sibelius Academy, Finland
mkuuskan@siba.fi

Abstract. In this paper we present a novel PWGL user-library, SDIF, that allows us to import SDIF encoded sound analysis information into PWGL. The library provides us with a visual box interface to the CNMAT/IRCAM SDIF Tools. The PWGL SDIF library builds on top of another PWGL library called SHELL enabling us to interact with the UNIX command-line. The SDIF library is still in early development and testing but it has already proven to be quite robust and functional. In this paper we introduce the current functionality of our library and discuss some concrete use cases and future development possibilities.

Keywords: SDIF, computer assisted composition, visual programming, visualization

1 Introduction

PWGL [9] is a Lisp-based music programming environment designed for the applications of computer assisted composition, music theory and analysis, software synthesis, and music notation. Currently, PWGL lacks tools that would allow us to straightforwardly use and manipulate sound analysis information in the visual patch. There are numerous free and commercial applications, such as AudioSculpt and SPEAR, that are able to perform sophisticated sound analysis and processing. The solution we propose here is to use third party applications to perform the analysis/processing and read the information into PWGL for further treatment. To this end we have implemented a new PWGL user-library, SDIF, that is able to import SDIF description files into our system.

SDIF (Sound Description Interchange Format, [12]) is a standard and extensible interchange format of sound descriptions jointly developed by IRCAM and CNMAT. SDIF consists of a large collection of spectral description types. Numerous programs currently support SDIF, among others Matlab [13], Max/MSP, PureData, jMax (through the FTM library), OpenMusic [2], Spear [6], AudioSculpt [4], ASAnnotation [3], and CLAM [1]. It has become de facto standard sound interchange format in the field of computer music and research.

In addition to the standalone applications several language bindings also exist, among others for C, Java, and Python. The Lisp-based SDIF interface is

* The work of Mika Kuuskankare has been supported by the Academy of Finland (SA137619). We would also like to thank CCRMA, Stanford University, for hosting the research.

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

provided by OpenMusic. In OpenMusic the SDIF interface [5] is implemented by calling the functions of a dynamic C-library directly using the Lisp Foreign language Interface (FLI). The SDIF data types as mirrored in the Lisp side using FLI data types.

Our interface, in contrast, provides direct access to the IRCAM/CNMAT SDIF Tools–a collection of command-line programs that help in reading, writing, and manipulating SDIF data files. Our library is implemented using a standard PWGL library called SHELL that allows us to interface with the UNIX command-line. SDIF operations are performed by the SDIF Tools and the results are read in PWGL through a pipe opened between PWGL and the shell. There are several advantages in this approach. First, it provides full access to the information distributed in SDIF format. Everything that can be done with the SDIF Tools can be done inside PWGL. Second, our box interface mirrors the functionality of the SDIF Tools. Thus, there's no need to learn a new box representation or naming conventions.

Eventually, the SDIF interface will allow us to develop many new PWGL applications. Our music notation program, ENP [7], in conjunction with the SDIF library, provides the users with a powerful toolkit, for example, for the applications of spectral composition. Our software synthesizer, PWGLSynth [10], would benefit from sophisticated sound analysis data difficult to obtain in any other way.

The PWGL SDIF user-library and the relating documentation can be downloaded from our project's web site at `http://www2.siba.fi/PWGL/downloads.html`

2 Overview of the SDIF Library

The PWGL SDIF library interfaces with the open-source CNMAT/IRCAM SDIF command-line tools. When the user loads our library for the first time, a PWGL specific version of the SDIF Tools is automatically compiled[1] and installed locally inside the SDIF library folder. Our version converts the output into a format understood by the Lisp reader, i.e, the results are in general always returned as a Lisp list.

The main utilities that we currently use are: `querysdif`, and `sdifextract`. `querysdif` displays a summary of the data in an SDIF file, and the ASCII header information. `sdifextract`, in turn, outputs the data of an SDIF file either as a whole or according to the options defined by the user.

The SDIF user-library is accompanied by a tutorial that gives several working examples dealing with compositional and sound synthesis applications. The documentation also provides the users with links to relevant study material that can be found either inside PWGL or on the internet.

Most of the functionality presented in the sections 2.1–2.3 the SDIF library inherits form the SHELL library. The PWGL SHELL library provides a collection

[1] The prerequisite for using the library is that the user has installed the free Apple Developer Tools.

of specialized box types that allow us to interface visually with the UNIX shell. Thus, it is possible to call virtually any shell program, to redirect and pipe commands, and to input the results back into the PWGL patch.

2.1 Managing Options

Command-line utilities usually allow users to define a variable number of different options. Managing and remembering all the possible options and the combinations thereof can be quite demanding. Our system provides the users with a visual browser (see Figure 1) for managing options. From the browser the user is able to select an appropriate option and study its documentation and even see concrete examples of its use. The SDIF library boxes behave like expert boxes in the sense that they provide relevant information for the users and assistance in their use.

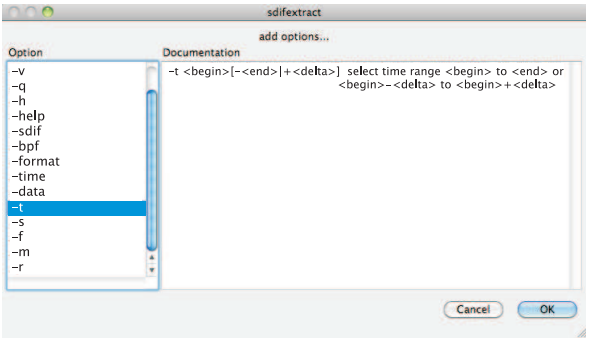


Fig. 1: *The SDIF library options dialog showing the options relevant to the sdifextract box.*

2.2 Error Handling

One important functionality of the SDIF library is its ability to gracefully handle errors that happened during the shell execution and to signal the user that something has gone wrong, e.g., a required file was not found. The UNIX error code is trapped by the boxes and displayed both visually and textually in the patch. The boxes can thus display pertinent information about their status and the success of the operation in question.

Figure 2 shows a box of the SDIF library in an erroneous state. A red warning sign is drawn over the box. When the user moves the mouse pointer over the box the relevant error message (corresponding to the UNIX error code) is displayed in a message area at the bottom of the patch window.

2.3 Argument Passing and Conversion

When interfacing with the SDIF Tools we need to be able to convert Lisp arguments into a form understood by the command-line utilities. The SDIF library

Fig. 2: *An SDIF library box showing that an error occurred while processing the user's request.*

provides a method with which the users can define translators from data written in Lisp to the data required by the utility.

Let us examine a brief example. Some of the SDIF tools accept time range as an option. For example, sdifextract allows us to define a time range for extracting only the relevant part of the audio analysis. This information is given using one of the following formats: `1.5-3.5` or `2.5+3.0`.

Normally, when passing this option from Lisp the users would have to write the data as a string, which makes its use inconvenient. In Lisp, the most convenient way of representing this information is a list of two numbers, e.g., '(1.0 2.5). The problem can be solved by adding a pre-processor to the SDIF box. In this case, we define a pre-processor for the given box type, `sdifextract`, and for a given option `-t` (as for time) as follows:

```
1 (add-pwgl-shell-box-pre-process
2 "sdifextract"
3 "-t"
4 #'(lambda(x)
5     (format () "~f-~f" (car x) (cadr x))))
```

The translator function (lines d-e) converts the incoming data. Now, the user can input the time range just by passing a list of two numbers.

3 SDIF Library Boxes

Here, we present the new PWGL box types (in addition to the two core boxes sdifextract and querysdif) that are unique to the SDIF library: (1) SDIF-selection box (2) SDIF-selection-spec box, (3) SDIF-range box, and (4) SDIF-type box.

3.1 SDIF-selection Box

Routinely there is a need to access only a subset of the information presented in an SDIF file. All SDIF tools accept an SDIF selection, which allows the users to specify relevant time ranges, frames or matrices that are accessed from the file. In our case this allows for a fast access to a part of the data contained by an SDIF file. Furthermore, the files do not need to be processed inside PWGL using Lisp; the processing is done by the SDIF library itself. This reduces the workload in PWGL and also allows us to remain compatible with the SDIF Tools collection as the implementation is kept outside our system. Incidentally, our flexible box

design scheme should allow us to keep up with the potential changes in the SDIF selection syntax.

The syntax of the selection is specified in [11] and is as follows:

```
[filename]::[#stream][:frame]
[/matrix][.column][_row][@time]
```

For convenience we have defined a special PWGL box (see Figure 3) that allows us to define the SDIF selection using Lisp objects. The box accepts a variable number of arguments in an arbitrary order. It sorts its arguments and outputs a syntactically valid SDIF selection object that can be passed as an argument to relevant SDIF library boxes.

Fig. 3: *An SDIF-selection box with two options: frame and time.*

The box with the options shown in Figure 3 translates to the following piece of SDIF selection code:

```
[...]/pf-w.pd-hit-fnl.fft.sdif::1HRM@0.0-2.0
```

3.2 SDIF-selection-spec Box

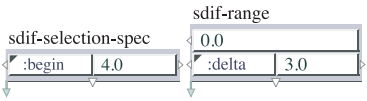
SDIF-selection-spec (Figure 4a) box converts lisp representations to various other formats required by the SDIF Tools. This box can be used with the SDIF-selection box to define ranges, or comma separated lists when needed. In SDIF, numeric values can be represented either as lists (e.g., selecting only the columns 1 and 2) or ranges (e.g., selecting the rows 1-50). The box also allows us to specify incomplete ranges as per SDIF specification (e.g., where the lower or upper value is replaced by the respective minimum or maximum value). The box shown in Figure 4a translates to an (incomplete) SDIF range specification `4.0-` (i.e., beginning from 4.0 seconds until the end).

3.3 SDIF-range Box

SDIF-range (Figure 4b) box converts lisp representations to the SDIF time range format. This box accepts the input in several different formats as per SDIF specification. The box shown in Figure 4b translates to a range of `0.0+3.0`.

3.4 SDIF-type Box

The SDIF Type box provides the users with a browsable dialog containing all the SDIF types and any relevant information about them. The box can be connected to other SDIF library boxes that require SDIF types as parameters. Multiple selection is also allowed.



(a) (b)

Fig. 4: *The SDIF library utility boxes: (a) the SDIF-selection-spec box, and (b) the SDIF-range box.*

4 Some Use Cases

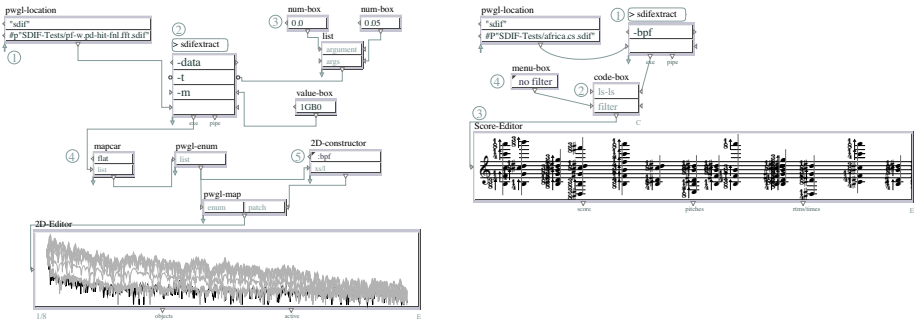
In this section we give two use cases of the SDIF library. In our first example, we visualize FFT data stored in an SDIF file. The second example deals with a compositional application where we read into PWGL chord sequence analysis information, prepared with the help of AudioSculpt, and convert it into a musical score.

4.1 Data Visualization

Figure 5a shows a patch where we read FFT information stored in an SDIF file and visualize it using our 2D-Editor [8]. The SDIF file pathname is given in (1) as an argument to the `sdifextract` box in (2) which, besides the mandatory pathname argument, has three options. The `-data` option instructs the box to return only the data without times. The `-t` option gets as an argument a time range, i.e., we read in only the frames between 0 and 5 milliseconds. The `-m` option allows us to select which matrix to extract. In this case, we are interested in matrices of the type `1GB0` which contain the FFT data. As the 2D-Editors can display information in several independent layers we use the 2D-constructor box (5) to create individual breakpoint-functions of each of the FFT-frames. The result is shown in the 2D-Editor at the bottom of the patch.

4.2 Score Interface

Figure 5b demonstrates how to manipulate SDIF data in PWGL to generate a sequence of chords. The analysis data is read in (1) using the `-bpf` option as we need the frame times as well as the frequency data. In (2) we convert in the code-box (the code is not shown here) the data to a list of chords and the result is given to the Score-Editor box (3). Here, the user can apply a filter (4) to get all notes, only notes that have velocity values below the average velocity of the chord, or only notes that have velocity values that are above average. Note also, that ENP allows us to represent micro tones using a dynamic resolution. Here, an eighth-tone resolution is defined by the user.



(a) (b)

Fig. 5: *The examples demonstrating the usage of PWGL SDIF user-library: (a) visualizing FFT information stored in an SDIF file, and (b) manipulating FFT data to generate symbolic music notation.*

5 Future Development

Currently, it is only possible to read into PWGL analysis information prepared outside our system. Next, we plan to extent the current interface scheme so that it supports making the analysis itself using the patch language. The extension would provide us with an access to SuperVP or pm2 (two sound processing tools developed by the Analysis/Synthesis team of IRCAM) or other similar analysis tools.[2] Furthermore, it should eventually be possible to write SDIF files directly from PWGL. This would make it possible to process the SDIF information in a patch and save the processed material again in SDIF format.

6 Conclusions

In this paper we present a new PWGL user-library that allows us to access and manipulate different kinds of sound analysis information visually by interfacing with the SDIF file format. The library does not use foreign language bindings, but instead interfaces with the UNIX command-line tools. We describe the current state of the library, present the basic concepts and the functionality of the visual box interface. We also cover some of the potential applications by demonstrating how to visualize SDIF data and convert it to high-level musical score representation.

[2] The possibility of using AudioSculpt kernels would naturally be available only for users that have a licensed copy of the software.

References

1. Amatriain, X., Arumí, P.: Developing cross-platform audio and music applications with the clam framework. In: Proceedings of the International Computer Music Conference. pp. 403–410 (2005)
2. Assayag, G., Rueda, C., Laurson, M., Agon, C., Delerue, O.: Computer Assisted Composition at IRCAM: From PatchWork to OpenMusic. Computer Music Journal 23(3), 59–72 (Fall 1999)
3. Bogaards, N., Yeh, C., Burred, J.J.: Introducing asannotation: a tool for sound analysis and annotation. In: Proceedings of International Computer Music Conference (2008)
4. Bogaards, N., Röbel, A., Rodet, X.: Sound analysis and processing with audiosculpt 2. In: Proceedings of International Computer Music Conference. Miami, USA (2004)
5. Bresson, J.: Sound processing in openmusic. In: Proceedings of the 9th International Conference on Digital Audio Effects – DAFx-06. pp. 325–330. Montréal, Canada (2006)
6. Klingbeil, M.: Software for spectral analysis, editing, and synthesis. In: Proceedings of the International Computer Music Conference (2005)
7. Kuuskankare, M., Laurson, M.: Expressive Notation Package. Computer Music Journal 30(4), 67–79 (2006)
8. Laurson, M., Kuuskankare, M.: PWGL Editors: 2D-Editor as a Case Study. In: Sound and Music Computing '04. Paris, France (2004)
9. Laurson, M., Kuuskankare, M., Norilo, V.: An Overview of PWGL, a Visual Programming Environment for Music. Computer Music Journal 33(1), 19–31 (2009)
10. Laurson, M., Norilo, V., Kuuskankare, M.: PWGLSynth: A Visual Synthesis Language for Virtual Instrument Design and Control. Computer Music Journal 29(3), 29–41 (Fall 2005)
11. Schwarz, D., Wright, M.: Extensions and applications of the sdif sound description interchange format. In: In Proceedings of the International Computer Music Conference. pp. 481–484 (2000)
12. Wright, M., Chaudhary, A., Freed, A., Wessel, D., Rodet, X., Virolle, D., Woehrmann, R., Serra, X.: New applications of the sound description interchange format. In: International Computer Music Conference. pp. 276–279. International Computer Music Association, Ann Arbor, Michigan (1998), `http://cnmat.berkeley.edu/publications/new_applications_sound_description_interchange_format`
13. Wright, M., III, J.O.S.: Open-source matlab tools for interpolation of sdif sinusoidal synthesis parameters. In: Proceedings of International Computer Music Conference. pp. 632–635 (2005)

Application of Pulsed Melodic Affective Processing to Stock Market Algorithmic Trading and Analysis

Alexis Kirke[1] and Eduardo Miranda[1]

[1] Interdisciplinary Centre for Computer Music Research, School of Humanities, Music and Performing Arts, Faculty of Arts, University of Plymouth,
Drake Circus, Plymouth, UK
{Alexis.Kirke, Eduardo.Miranda}@Plymouth.ac.uk

Abstract. The application of Pulsed Melodic Affective Processing (PMAP) to stock market analysis and algorithmic trading is examined. PMAP utilizes musically-based pulse sets ("melodies") for processing – capable of representing affective states. Affective processing and affective input/output is now considered to be a key tool in artificial intelligence and computing. However in the designing of processing elements (e.g. bits, bytes, floats, etc), engineers have primarily focused on the processing efficiency and power. Having defined these elements, they then go on to investigate ways of making them perceivable by the user/engineer. But the extremely active and productive area of Human-Computer Interaction - and the increasing complexity and pervasiveness of computation in our daily lives – supports the idea of a complementary approach in which computational efficiency and power are more balanced with understandability to the user/engineer. PMAP provides the potential for a person to tap into the affective processing path to hear a sample of what is going on in that computation, as well as providing a simpler way to interface with affective input/output systems. In this paper PMAP will be applied to a simple algorithmic trading system based on an affective model of a simulated stock market.

Keywords: Human-Computer Interaction, Music, Affective, Boolean Logic, Neural Networks, Behavioural Finance, Algorithmic Trading, Stock Market

1 Introduction

Pulsed melodic affective processing involves the use of music as a processing tool for affective computation in artificial systems. It has been shown that affective states (emotions) play a vital role in human cognitive processing and expression [1]. The field of Behavioral finance has highlighted the importance of emotions in finance and markets [2]. This paper proposes the application of PMAP in stock market trading and analysis.

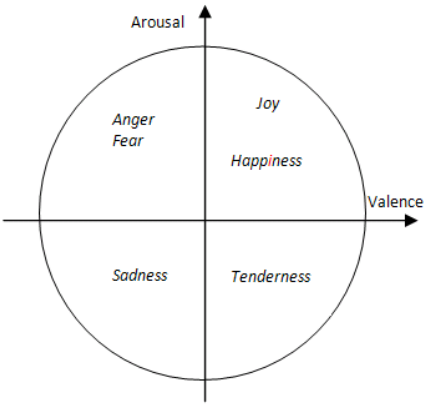
Affective state processing has been incorporated into previous artificial intelligence processing and robotics [3]. The issue of developing systems with affective intelligence which also provide for greater user-transparency, is a key

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

element discussed in this paper. Music has often been described as a language of emotions [4]. There has been work into automated systems which communicate emotions through music [5] and which detect emotion embedded in music based on musical features [6]. Hence the general features which express emotion in western music are known.

Before introducing these, affective representation will be briefly discussed. The dimensional approach to specifying emotion utilizes an n-dimensional space made up of emotion "factors". Any emotion can be plotted as some combination of these factors. For example, in many emotional music systems [7] two dimensions are used: Valence and Arousal. In that model, emotions are plotted on a graph (see Figure 1) with the first dimension being how positive or negative the emotion is (Valence), and the second dimension being how intense the physical arousal of the emotion is (Arousal). For example "Happy" is high valence high arousal affective state, and "Stressed" is low valence high arousal state.

Figure 1: The Valence/Arousal Model of Emotion



Previous research [8] has suggested that a main indicator of valence is musical key mode. A major key mode implies higher valence, minor key mode implies lower valence. For example the overture of The Marriage of Figaro opera by Mozart is in a major key; whereas Beethoven's melancholy "Moonlight" Sonata movement is in a minor key. It has also been shown that tempo is a prime indicator of arousal, with high tempo indicating higher arousal, and low tempo - low arousal. For example: compare Mozart's fast overture above with Debussy's major key but low tempo opening to "Girl with the Flaxen Hair". The Debussy piano-piece opening has a relaxed feel – i.e. a low arousal despite a high valence.

Affective Computing [9] focuses on robot/computer affective input/output. Whereas an additional aim of PMAP is to develop data streams that represent such affective states, and use these representations to internally process data and compute actions. The other aim of PMAP is more related to Picard's work – to aid easier sonification of affective processing [10] for transparency in HCI, i.e. representing non-musical data in musical form to aid its understanding. Related sonification

research has included tools for using music to debug programs [11]. There have also been a number of papers published on sonification [12][13] of stock market data, but not from an affective point of view, nor with a unified data stream for processing and sonification.

2 PMAP Representation of Affective State

Pulsed melodic affective processing is a method of representing affective state using music. In PMAP the data stream representing affective state is a series of pulses of 10 different levels with a varied pulse rate. This rate is called the "Tempo". The pulse levels can vary across 12 values. The important values are: 1,3,4,5,6,8,9,10,11,12 (for pitches C,D,Eb,E,F,G,Ab,A,Bb,B – we assume all melodies are in C). These values represent a valence (positivity or negativity of emotion). Values 4, 9 & 11 represent negative valence (Eb, Ab, Bb are part of C minor) e.g. sad; and values 5, 10,& 12 represent positive valence (E, A, B are part of C major), e.g. happy. The other pitches are taken to be valence-neutral. For example a PMAP stream of say [1,1,4,4,2,4,4,5,8,9] (which translates as C,C,Eb,Eb,C#,Eb,Eb,E,G,Ab) would be principally negative valence since most of the notes are in the minor key of C. It is understood that these key modes can become ambiguous, particularly in terms of relative keys. It is expected that they will tend to be more accurate for more extreme valence values (i.e. those furthest from 0).

The pulse rate of a stream contains information about arousal. So [1,1,4,4,2,4,4,5,8,9] transmitted at maximum pulse rate, could represent maximum arousal and low valence, e.g. "Anger". Similarly [10,8,8,1,2,5,3,1] (which translates as A,G,G,C,D,E,C,C) transmitted at a quarter of the maximum pulse rate could be a positive valence, low arousal stream, e.g. "Relaxed" (because it is in the major key of C). If there are two modules or elements both with the same affective state, the different note groups which go together to make up that state representation can be unique to the object generating them. This allows other objects, and human listeners, to identify where the affective data is coming from.

In performing some of the analysis on PMAP, it is convenient to utilize a parametric form, rather than the data stream form. The parametric form represents a stream by a Tempo-value variable and a Key-value variable, which can vary continuously as Arousal and Valence vary.

3 Affective Market Mapping

PMAP has been applied to affective processing in a multi-robot security system, and to a text emotion detection system [14]. The first of these was achieved by designed a set of "musical" logic circuits. The second involved a neural network based on musical neurons or "Murons", whose weights adjust tempo and key mode, and which learned by gradient descent. There are 3 elements which suggest PMAP may have potential in the stock markets: a simple market-state mapping (described below), the incorporation of trader, client and news article "sentiment" into what is an art as well as a science, and a natural sonification for eyes-free HCI in busy environments. The Affective Market Mapping (AMM) involves mapping stock movements onto a PMAP representation. Such a mapping would allow PMAP processing to interact with stock market data and be used for algorithmic trading. One mapping that was initially considered was a risk / return mapping – letting risk be

mapped onto arousal / tempo, and return be mapped onto valence / key mode. However this does not give an intuitively helpful result. For example it implies that a high arousal high valence stock (high risk / high return) is "happy". However this entirely depends on the risk profile of the investor / trader. So a more flexible approach – and one that is simpler to implement - for the AMM is:

1. Key mode (valence) is proportional to Market Imbalance.
2. Tempo (arousal) is proportional to Number of Trades per Second.

These can refer to a single stock, a group of stocks, or a whole index. Consider a single stock S. The Market Imbalance Z in a time period dT is the total number of shares of buying interest in the market during dT minus the total number of shares of selling interest during dT. This information is not publically available, but can be approximated. For example it can be calculated as in [15] - the total number of buy-initiated sales minus the total number of sell-initiated trades (normalized by the Average Daily Volume for S); with a trade is defined as buy initiated if it happens on an uptick in the market price of stock S, and sell-initiated if it happens on a downtick (the "tick algorithm"). If there are as many buyers as sellers in stock S then it is balanced and its market imbalance Z will be 0. If there are a large number of buyers and not enough sellers (e.g. in the case where positive news has been released about the stock) the imbalance will become positive.

To generate a melody from a stock, simply have a default stream of non-key notes at a constant or uniformly random rate; and every time there is a trade add a major key note for a buy initiated trade and a minor key note for a sell initiated trade. So for example, if a stock is being sold off rapidly due to bad news, it will have a negative market imbalance and a high trading rate – which will be represented in PMAP as a minor key and high tempo – earlier labelled as "angry". Stocks trading up rapidly on good news will be "happy", stocks trading up slowly in a generally positive market will be "relaxed". The resulting PMAP stream matches what many would consider their affective view of the stock.

6 Simulation

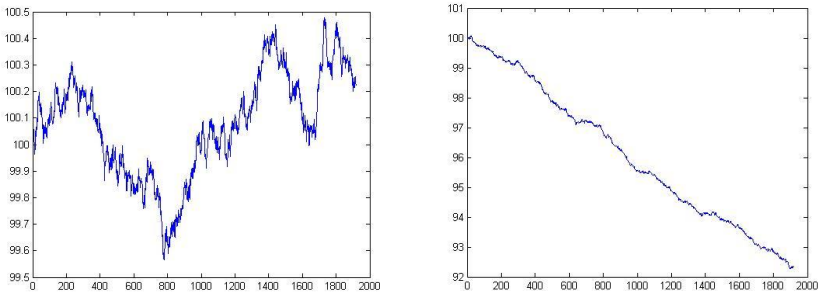
To examine a simple processing usage of the Affective Market Mapping and PMAP a basic algorithmic trading system will be implemented. Algorithmic Trading has become extremely prominent in the markets in the last few years [16], but we are not aware of any work which focuses on affectivity. To examine this approach, a simple stock market order book simulation has been developed. The market contains a single stock whose initial price is $100. Orders arrive at the market at a constant rate of one every 10 minutes. The stock has an Average Daily Volume of around 40000 shares. Each trade can be a buy or sell order with a probability p of being a buy order and 1-p of being a sell-order. The order book can contain up to 30 buy orders and 30 sell orders. Each order is uniformly randomly sized. The market price $p(t)$ evolves based on whether an order is a buy or sell order, the order size, and a price volatility parameter.

$$p(t) = p(t-1) + priceDriftFactor.orderSize.orderPrice / ADV - volatility + 2.r.volatility$$

The level at which a simulated order is priced is the market price *p(t)* with a certain deviation of % size defined by a parameter *priceFluctFactor*. Once the book has filled up with arriving orders, new orders overwrite the oldest ones. Although in the simulation it is known precisely whether the order is a buy or sell order, the tick algorithm is still used to estimate the order side for the affective market mapping. The accuracy of this estimation will depend on the size of the random price fluctuations in orders and the market price volatility – i.e. the higher the volatility and fluctuation parameters, the less accuracy the tick algorithm with exhibit. For the simulation detailed here *volatility* was set to 0.02, *priceDriftFactor* to 0.005 and *priceFluctFactor* to 0.001. This led to the tick algorithm being on average about 75% accurate. In other words about 75% of orders were correctly classified. If volatility is increased to 0.005, the accuracy drops to around 60%.

To see how this model functions with the affective marking mapping, consider the prices of a month's worth of trading shown on the left hand side of Figure 2, where maximum order size is 1000 shares. This month is a "neutral" month – in other words the probability of a buy order is equal to the probability of a sell order. The right hand side of Figure 2 shows a month where there is a constant probability of 70% of a sell order arriving, and of 30% of a buy order arriving. The left side of Figure 3 shows the valence calculated for this "selling month". The higher valences are equivalent to a more clearly major key mode and the lower valences to a more clearly minor key mode. (Note in the following discussions valence and arousal are used interchangeably with key mode and tempo. This is for simplicity - rather than constructing a melodic stream.) The first thing to observe is that the valence is usually negative, with a mean valence of -0.34. There are 5 sections where it goes above 0, but then there are also local maxima in the globally falling stock price in Figure 2.

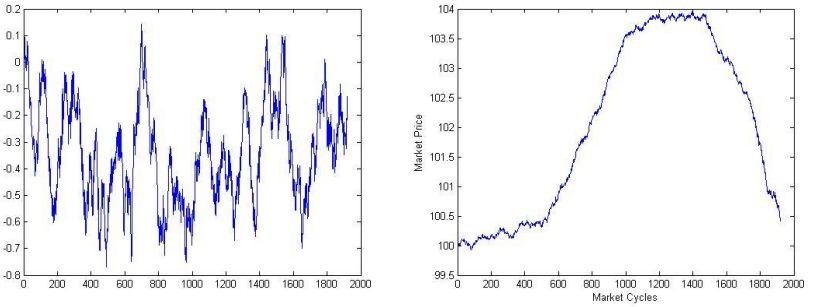
Figure 2: Stock Price in dollars in a "Neutral" Month; and in a "Selling Month" (x-axis is simulation time steps)



It is much clearer to see patterns of behaviour if both key-mode and tempo are plotted as valence and arousal, as in Figure 4. The right hand of Figure 3 shows a market event which begins with a relative relaxed trading in the stock just above $100, followed by a rapid rise in the stock price due to an increase in buy order probability. This is followed by another period of stable price trading just below $104, then for some reason the stock starts to fall with increasing rapidity back to just above $100. This is done by setting the buy probabilities to 0.5, 0.75, 0.5, 0.25 respectively;

and setting average order amounts to 1000, 2000, 1500, and then during the selling period to 1500 and then 4000. Looking now at how this is reflected in the affective market model, we can observe the left hand of Figure 4. To clarify this further an average version is shown in the right hand side of Figure 4, averaged over 50 runs.

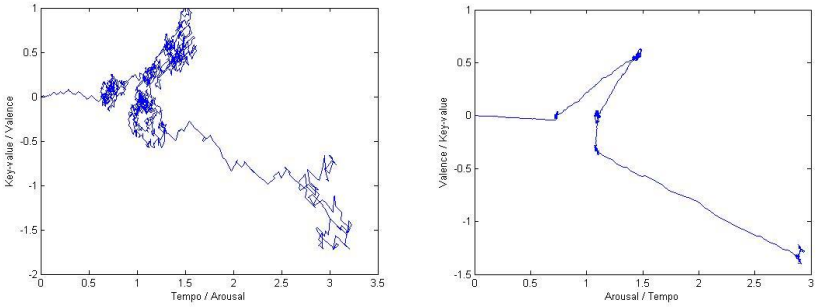
Figure 3: Valence of Stock Price in the "Selling" Month; and Price during "Event" (x-axis is simulation time steps)



The stock begins at the far left of the diagram with a low arousal and neutral valence due to the slow build of the order book (which starts from empty). One can then observe at least 5 "emotional regimes" that the market moves through, as the arousal/valence line is followed by eye moving from the far left to the far right of the diagram:

1. "Relaxed" – after the arousal builds up there is a regime around 0.02 arousal at the left of the diagram.
2. "Joyful"/"Excited" – this is the region of maximum valence / key-value and with significantly increase arousal / tempo, during which the stock price is rising more rapidly.
3. "Happy" – The market rise is slowing down as it approaches $104
4. "Sad" – The market starts to go down slowly.
5. "Angry"/"Fearful" – At around $102.50 the stock begins to fall rapidly.

Figure 4: Affective Market Model of Stock event; and Mean Affective Market Model averaged over 50 Stock events



An interesting element to observe concerning these regimes is that they are audible since if sound is played with the relevant key-value and tempo the music will – for western listeners - have the affective communication (approximately) of: “Relaxed”, “Excited”, “Happy”, “Sad”, and “Fearful” [5].

To examine how the AMM might be used in algorithmic trading, consider a simple rule:

If keymode > trigger then buy stock quantity proportional to tempo

If keymode <-trigger then sell short the stock with quantity proportional to tempo

Using this rule and the above market model, with a trigger value of 0.1, trading simulations were run (once again substituting valence for key-mode, and arousal for tempo). When the trigger kicked in a stock quantity of 50xArousal was traded. So an arousal of 0 would lead to a trade of 0 shares, an arousal of 2 would lead to a trade of 100 shares. The results are shown in Table 1, each cell gives the average profit from 50 experiments.

Table 1: Profits for the Strategy

Trigger	Arousal-based Trade Size?	Trigger Strategy Profit	Random Strategy Profit
0.6	Y	$4,515	$671
0.6	N	$9,109	-$244
0.4	Y	$21,618	$17
0.4	N	$18,333	$76
0.1	Y	$15,609	$1,843
0.1	N	$18,235	-$103

The Random strategy trades with approximately the same frequency as the Trigger Strategy but at randomize times and random order sizes. Column 2 is included so as to compare trading a fixed amount, with trading an amount decided by arousal level. It can be seen that the trigger strategy outperforms the random strategy, and that a full valence / arousal strategy (where trade size is based on arousal) outperforms a valence-only strategy. Another interesting element of the arousal-based order size is that order sizes will tend to be closer to the immediate market volumes, which may tend to reduce transaction costs. Note that algorithmic strategies such as the above could be embedded in Music Logic circuits and Musical Neural Networks [14], allowing them to interact with other PMAP functionality such as sentiment analysis of news text feeds.

In theory the above stock market methodologies could all have been derived purely based on valence and arousal, without mentioning tempo and key mode. However PMAP is designed to simplify the sonification of internal processing [14]. So this work is designed to show another area where PMAP can be applied, rather than to address specifically how the sonification of internal processing has particular benefits in stock market computations. But there is also a benefit which stands out here in the use of PMAP – it incorporates a sonification of the market. The melodies provide a natural sonification of stock movements – a useful factor for traders whose eyes are already too busy [13]. One can also consider the harmonic relationship between two stocks, or between a stock and the market. There may be PMAP methods developable

such that if stocks start to create cross-dissonance where once was consonance (e.g. one becomes more major as the other stays minor) then this indicates a potential divergence in any correlated behaviour.

8 Conclusions

This paper has introduced the concept of PMAP and given an initial proof of concept of a stock market application. PMAP is a complementary approach in which computational efficiency and power are more balanced with understandability to humans (HCI); and which can naturally address rhythmic and affective processing. Its application in stock markets requires significant further work before these very initial experiments. For example, it needs to be tested against real stock market data, and trading profits need to include the factoring in of transaction costs. There also need to be cross-comparisons with other standard algorithmic trading approaches. Finally there should be listening tests analysing the audible information content in the processing and the market mapping.

There are a significant number of issues to be further addressed with PMAP itself, a key one being is the rebalance between efficiency and understanding useful and practical, and also just how practical is sonification? The valence/arousal coding provides simplicity, but is it sufficiently expressive while remaining simple? Similarly it needs to be considered if a different representation than tempo/key mode be better for processing or transparency. PMAP also has a close relationship to Fuzzy Logic and Spiking Neural Networks – so perhaps it can adapted based on lessons learned in these disciplines. And finally, most low level processing is done in hardware – so issues of how PMAP hardware is built need to be investigated.

References

1. Malatesa, L., Karpouzis, K., Raouzaiou, A.: Affective intelligence: the human face of AI, In Artificial intelligence, Springer-Verlag (2009).
2. Davies, G.B, De Servigny, A.: Behavioral Investment Management: An Efficient Alternative to Modern Portfolio Theory, McGraw-Hill (2012).
3. Banik, S., Watanabe, K., Habib, M., Izumi, K.: Affection Based Multi-robot Team Work, In Lecture Notes in Electrical Engineering, pp. 355--375 (2008).
4. Cooke, D.: The Language of Music. Oxford University Press (1959).
5. Livingstone, S.R., Muhlberger, R., Brown, A.R., Loch, A.: Controlling Musical Emotionality: An Affective Computational Architecture for Influencing Musical Emotions. Digital Creativity, Vol. 18, No. 1, pp. 43--53 (2007).
6. Kirke, A., Miranda, E.: Emergent construction of melodic pitch and hierarchy through agents communicating emotion without melodic intelligence, In Proceedings of 2011 International Computer Music Conference (ICMC 2011), ICMA (2011).
7. Kirke, A., Miranda, E.: A Survey of Computer Systems for Expressive Music Performance, ACM Surveys, Vol. 42, No. 1, pp. 1--41 (2009).
8. Juslin, P.: From Mimesis to Catharsis: expression, perception and induction of emotion in music, In Music Communication, pp. 85--116, Oxford University Press (2005).
9. Picard, R.: Affective Computing: Challenges, International Journal of Human-Computer Studies, Vol. 59, No. 1-2, pp. 55--64 (2003).
10. Cohen, J.: Monitoring Background Activities, In Auditory Display: Sonification, Audification, and Auditory Interfaces, pp. 499--531 (1994).
11. Vickers, P., Alty, J.: Siren songs and swan songs debugging with music. Communications of the ACM, Vol. 46, No. 7, pp. 87--92 (2003).

12. Worral, D.: The use of sonic articulation in identifying correlation in capital market trading data, In Proceedings of the 15th International Conference on Auditory Display (2009).
13. Cifariello, F., At F.: sMAX: A Multimodal Toolkit for Stock Market Data Sonification, In Proceedings of the 10th International Conference on Auditory Display (2004).
14. Kirke, A., Miranda, E.: Pulsed Melodic Processing - Using Music for natural Affective Computation and increased Processing Transparency, In Music and Human-Computer Interaction, Springer Verlag (2012).
15. Kissell, R., Glantz, M.,: Optimal Trading Strategies, Amacom (2003).
16. Pole, A.: Algorithmic Trading and Statistical Arbitrage. Trading, Vol. 2, No. 1, pp. 79—87 (2007).

A Graph-Based Method for Playlist Generation

Debora C. Correa[1], Alexandre L. M. Levada[2] and Luciano da F. Costa[1]

[1] Instituto de Fisica de Sao Carlos, Universidade de Sao Paulo, Sao Carlos, SP, Brazil
[2] Departamento de Computacao, Universidade Federal de Sao Carlos, SP, Brazil
deboracorrea@ursa.ifsc.usp.br, alexandre@dc.ufscar.br,
luciano@ifsc.usp.br

Abstract. The advance of online music libraries has increased the importance of recommendation systems. The task of automatic playlist generation naturally arises as an interesting approach to this problem. Most of existing applications use some similarity criterion between the songs or are based on manual user interaction. In this work, we propose a novel algorithm for automatic playlist generation based on paths in Minimum Spanning Trees (MST's) of music networks. A motivation is to incorporate the relationship between music genres and expression of emotions by capturing the presence of temporal rhythmic patterns. One of the major advantages of the proposed method is the use of edge weights in the searching process (maximizing the similarity between subsequent songs), while Breadth-First (BF) and Depth-First (DF) search algorithms assume the hypothesis that all the songs are equidistant.

Keywords: playlist, rhythm, graphs, search algorithms.

1 Introduction

With the dissemination of online resources with music content, music recommendation systems have received much attention. Indeed, the sometimes manual and time-consuming selection task of music playlists can be replaced by automatic algorithms. Such algorithms can generate the playlist according to the user's music preferences or through some defined similarity criterion between the songs.

We can relate three main important aspects of a playlist: the individual songs themselves, the order in which they are played, and the size of the playlist. In the literature, we can find earlier efforts concerning the automatic generation of musical playlists [1, 10, 7, 9]. Most of these approaches are based on collaborative filtering techniques, audio content analysis, and require a manually labeled database or the analysis of metadata.

In this scenario, our contribution is to propose a novel and unsupervised playlist algorithm, avoiding possible noise due to the user's subjectiveness. Besides, a motivation is the possibility to associate music genres to the presence of temporal patterns in the rhythm as a way to express notions of emotion. According to [12], regular and smooth rhythmic patterns indicate expressions like happiness, joyness or peacefulness. Irregular and complex rhythms indicate

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

expressions like amusement, tension or uneasiness, while fluent rhythms indicate feelings like happiness, gracefulness or dreamy. Thus, in our playlists the songs are rhythmically related and, therefore, emotionally linked.

The algorithm is performed on a Minimum Spanning Tree (MST) extracted from music networks. Although we have not performed a conclusive user evaluation of the subjective quality of the the playlists, this investigative work relies on the properties of the MST, reflecting the overall characteristics of the playlists. The similarity criterion used to build the music networks is based on the cosine distance between feature vectors of the songs. Thus, the playlist can be easily adapted to other types of musical feature and similarity metrics [13].

The remaining parts of the paper are organized as follows: section 2 describes the database, the methodology to construct the music networks, and the extraction process of the note duration dynamics; section 2.4 presents the proposed algorithm for automatic generation of the music playlists and a discussion of its characteristics. Finally, section 3 contains the main conclusions.

2 Constructing music graphs

2.1 The database

Our database consists of four musical genres with seventy samples each: blues, *mpb* (*Brazilian popular music*), reggae and rock. Our motivation for choosing these four genres is the availability of MIDI samples in the Internet with considerable quality and the different tendencies they represent.

MIDI format is simpler to analyse than audio files, since all voices are separated in tracks. However, as it is a symbolic representation, MIDI allows a clear analysis of the involving music elements, in opposite to audio files in which all information is mixed together. To read a MIDI file, we used the Sibelius software and the free Midi Toolbox for Matlab computing environment [11]. This toolbox provides a note matrix representation, with information like relative duration (in beats), MIDI pitch, and others. The note value is represented in this matrix through relative numbers. To deal with possible fluctuations in tempo, we unset the "Live Playback" option in Sibelius. In this way, the note values in the MIDI file respect their relative proportion (e.g, the eighth note is always 0.5).

As we want to generate rhythm-based similarity playlists, we propose a similarity criterion based on the temporal sequence of the note values present in the percussion track. In this work, the instrumentation is not considered. If two or more note events occur in the same beat, the median duration of them is taken. To clarify this concept, Figure 1 shows the first measures of the percussion track of the music *Who can it be now?* (Men at Work).

Part of the matrix representation of the notes values of the third measure is presented in Table 1. Taking the median value for events occurring at the same beat, the final vector with note values is: [0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5]. The note vector of the whole percussion is computed for each song in the database. All this process can be performed automatically.

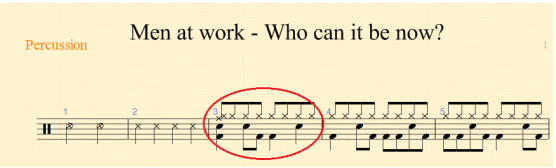


Fig. 1. Example of a percussion track.

Beat	8	8	8	8	8.5	9	9	9.5	9.5	10	10	10.5	11	11	11.5
Relative Duration	0.5	0.5	1	1	0.5	0.5	0.5	0.5	0.5	0.5	1	0.5	0.5	1	0.5

Table 1. Matrix representation of first measure of the percusion in Figure 1.

2.2 Markov modeling for note value dynamics

Markov chains establish a conditional probability structure in which the probability of an event n depends on one or more past events $(n-1, n-2...)$[5]. The order of the chain is determined by the number of past events taken into consideration. A first order Markov chain models the dependency between the current event and its predecessor. Similarly, a second order Markov chain takes into account the two subsequent past events. Therefore, a nth-order Markov chain will set up a transition matrix of $n+1$ dimensions, in which each entry gives the conditional probability of an event, based on its previous n states. We propose that the events to be modeled in the Markov chains be the note values extracted from the percussion track of the songs. For each note vector, we compute the first and second order probability transition matrices.

In order to reduce data dimensionality, we performed a preliminary analysis of the relative frequency of note values and pairs of note values concerning all the songs, in a way that extremely rare transitions were discarded[1]. We considered 18 different note values in the first order Markov chain (isolated note values) and 167 different pairs of note values for the second order Markov chain. Therefore, the first order probability transition matrix is 18 (rows) x 18 (columns), indicating the conditional probability that a note value i is followed by a note value j. The second order probability transition matrix is 167 (rows) x 18 (columns), indicating the conditional probability that a pair of note values represented in row i is followed by a note value in column j. Both matrices are treated as a 1 x 324 and 1 x 3006 (167 * 18) feature vector, respectively. To form the final feature vector of each song, we concatenate both feature vectors (3330 dimensions).

2.3 Music Networks and Playlist Generation

A complex network is a graph that exhibits non-trivial topological structure between its elements [6]. A graph can be represented by vertices, edges (links),

[1] We count how many times each note duration appears considering all songs and discard the ones that occurred less than 0.1% of times

and weights associated to the links. In this case, each edge has the form $w(i, j)$, expressing the weight $w(i, j)$ in the connection from vertex i to vertex j.

Each song is represented by a vertex. The weight of the link between two songs is expressed by the distance between their feature vectors. We considered the cosine distance, although many alternative distance metrics can be used [13]. However, it may be difficult to analyse intricate structures if the network is full-connected. Therefore, the motivation is to construct the playlist under the properties of a MST. Different distance metrics leads to different MSTs and, consequently, to different playlists. From the point of view of playlist generation, these variations are profit, allowing many possibilities of sequences of music.

The proposed music network allows a straightfoward automated recommendation scheme through a simple playlist generation algorithm, based on paths on a MST. Since the networks are built in a completely unsupervised manner, the user labeling is not required. The idea is to generate a playlist according to a similarity criterion, assured both by the network construction process, and by the MST properties (connectivity among all vertices and minimum dispersion).

Suppose one wants to perform a search for similar vertices in a music network. A possible solution is to apply a simple weighted random walk, a BF search algorithm (broadcast) or even a degree-based searching on the original graph [3]. However, in these situations, one may reach undesirable vertices once there is no way to assure that transitions between vertices are smooth in the sense that each visited vertex is somehow similar to the previous ones.

We propose that a MST can be used in the definition of a new search strategy, named the jumping walk algorithm. In few words, a MST is a minimum-weight acyclic connected subgraph in which there is only one single path from any vertex A to any other vertex B. Moreover, due to its properties, if vertex A is linked to vertices B and C by using a distance metric weighted edge, it means that both B and C are the two most similar ones regarding vertex A (considering the tree and not the original graph). Thus, visiting nearest neighbors in a MST will produce a sequence of vertices in which the next element is the most similar to the previous ones in the tree. We used the Prim's algorithm to generate a MST from the full-connected music network [8, 6].

2.4 The Jumping Walk Algorithm

The searching strategy is based on walks on MST's without repetition of vertices. A simple approach for playlist generation is to perform a breadth-first search in the resulting MST. The basic procedure is to form a list of songs by first visiting all vertices that are a distance one from the starting node, then visiting all vertices that are a distance two and so on until we reach a maximum depth. Although simple and functional, this approach has some drawbacks. First, it completely ignores the edge weights, assuming the hypothesis that all vertices are equidistant in the feature space, which is not reasonable. And second, the set of vertices that are located a distance d from a starting node are not linked in the MST, which means that these samples may not be close enough in the feature space. This effect is even worse for large values of d, where the equidistant

samples can be from totally different clusters. However, when walking through a MST, an unexpected situation may occur if we hit a leaf of the tree (dead-end path). The proposed strategy deals with such situation by jumping to the nearest unvisited vertex and restarting a new walk, as described in the following.

The jumping walk algorithm consists of a sequence of walks in the MST of a music network. The algorithm starts by choosing a random initial vertex representing a specific music genre, given as input by the user. After that, a sequence of vertices is defined by the continuous choice of the nearest unvisited neighbor (minimum weight edge) to be the next song. If the selected song is a leaf (that is, there is no unvisited neighbor), then we check all the vertices that are a distance $d = 2, 3, 4...$ until we find an unvisited vertex. In case of more than one option, we perform a jump to the nearest one according to the edge weights (this is equivalent to finding the shortest-path between the leaf and an unvisited vertex using the Djikstra algorithm). Even though in this case there is a jump, that is, a discontinuity in the sequence, the MST properties guarantees that the next song is the most similar song in the tree that was not played yet. Overall, after each jump (from a leaf of the tree), a new walk is started.

A potential problem with this method is the generation of playlists with M songs, where $M \approx N$ (N is the number of songs). With the rule of walking to the nearest neighbor (in terms of minimum weight), some branches of the MST may be neglected, which may cause large jumps as the number of unvisited vertices aproximates zero. Two ways to avoid this problem are: 1) having a sufficient large N (e.g., 100000); 2) stopping the walks whenever the number of visited vertices reaches an upper bound (e.g., $0.75 * N$).

Figure 2 illustrates the jump process. Suppose that we start in A, the sequence of visited vertices before we reach a leaf will be A, B and C. After that, a jump to the nearest vertex is performed. Note that the edges AB and BC are in the minimum path from C to D and also in the minimum path from C to E. Therefore, the next selected song (D in this case) will be the nearest to both C (position i in the list) and A (position $i - 2$ in the list).

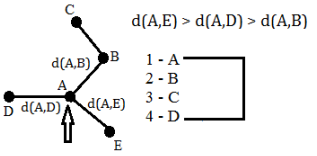


Fig. 2. An illustration of the jumping walk process.

Figure 3 shows the MST for the music network using the cosine distance between the vertices attributes. Many songs of the same genre in branches of the a same subtree. Applying the jumping walk algorithm and using the song number 39 as initial vertex, the top twenty five songs of the playlist are:

39 - *(blues)* Johnny Winter - Good morning little schoolgirl 50 - *(blues)* Stevie Ray Vaughan - Cold shot 62 - *(blues)* Stevie Ray Vaughan - Tell me 23 - *(blues)* Delmore Brothers - Blues stay away from me 203 - *(reggae)* Third World - Now that we've found love 25 - *(blues)* ElvesPresley - A mess of blues 44 - *(blues)* Ray Charles - Born to the Blue 12 - *(blues)* Barbra Streisand - Am I Blue? 38 - *(blues)* John Lee Hooker - One bourbon one scotch one beer 16 - *(blues)* Boy Williamson - Dont start me talking 18 - *(blues)* Boy Williamson - Keep it to yourself 152 - *(reggae)* Bob Marley - I shot the sheriff 36 - *(blues)* John Lee Hooker - Boom boom boom 3 - *(blues)* AlbertKing - Stormy monday 34 - *(blues)* Jimmie Cox - Before you accuse me 190 - *(reggae)* Natiruts - Liberdade pra dentro da cabeca 5 - *(blues)* BB King - Dont answer the door 193 - *(reggae)* Nazarite Skank 263 - *(rock)* Rolling Stones - Angie 239 - *(rock)* Metallica - Fuel 245 - *(rock)* Metallica - Sad but true 29 - *(blues)* Freddie King - Hide away 170 - *(reggae)* Cidade Negra - Eu fui eu fui 40 - *(blues)* Koko Taylor - Hey bartender 59 - *(blues)* Stevie Ray Vaughan - Manic depression

The resulting playlist is based on the temporal patterns of the note values in the percussion. Hence, the algorithm will not necessarily generate a sequence of songs belonging to the same genre, but, instead, a sequence in which the songs are similar according to the note value patterns, which can indicate a relationship between subject aspects such as mood and emotion.

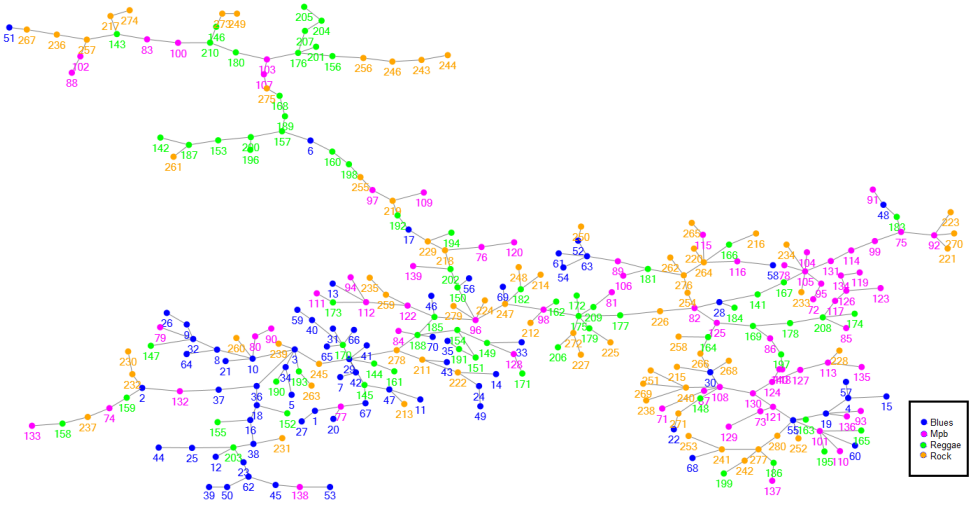


Fig. 3. MST for the network built with the cosine distance between vertices.

Different MSTs (derived from distincted music networks) lead to different playlists. In fact, multiple executions of PRIM's algorithm on the same network do not guarantee that the obtained MST's will be the same. For the purpose of playlists generation, this is a positive characteristic (controlled variability).

In a full-connected network, each song is connected to all other songs. If instead we link a song to its k nearest songs, we will have a k-regular network (digraph), since only the k nearest songs are linked together, avoiding long distance inter-genre connections. With the purpose of comparing the proposed approach for playlist generation with the BF search algorithm, we empirically

chose $k = 10$. In all experiments, when we refer to the BF search in the digraph, we mean that the search was performed in the 10-regular digraph.

For the first 100 songs in different playlists (using song 39 as the initial vertex), we analyzed the temporal aspects concerning the distance of subsequent songs as well as the inter-genre dynamics (Figure 4). We compared the JW algorithm with two different variations of the BF search algorithm: BF on the MST; BF on the 10-regular digraph. We can observe different behaviors depending on the adopting strategy. The mean distance is also presented for each situation. It shoud be noted that the JW algorithm produces the minimum mean inter songs distance. This means that, in average, the selected songs are more closely related than in the other methods. We analyzed the pairs of subsequent music genres for all the obtained playlists. The number of transitions between genres for each method were: JW = 57, BF on the MST = 60, BF on the 10-regular digraph = 73, indicating that the playlist produced by the JW algorithm was able to minimize the transitions between genres. Eventually, the distance between a pair of songs on a digraph may be undefined (as occured in Figure 4.)

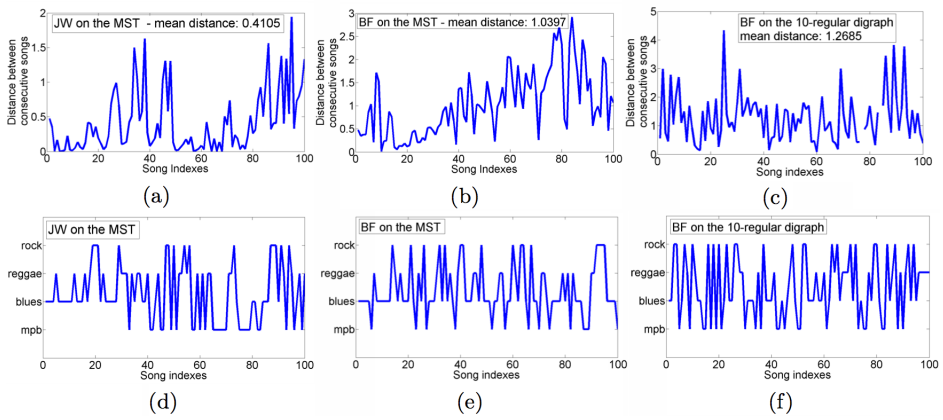


Fig. 4. Distances between subsequent songs and inter-genre interaction for the first 100 songs in the playlist obtained for the JW algorithm and BF searches.

3 Final remarks and ongoing work

We proposed a novel automatic algorithm for rhythm-based playlist generation based on the minimum spanning trees of music networks, It does not require user labeling, and can be easily implemented and adapted to other musical features.

In summary, the main contributions of the proposed method are: 1) the distance between consecutive songs in the playlist is minimized compared to BF search; 2) it reduces the number of abrupt transitions between genres; 3) songs from different genres are placed together if they have common rhythmic

patterns, which allows a controlled variability; 4) songs placed closely together also tend to have similar subjective aspects of mood.

Besides, BF searches on digraphs provide a much higher degree of freedom in the sense that we would face more unreliable inter-genre connections. However, we cannot guarantee the preference for the playlist obtained by MST paths or by BF searches on digraphs. It will mainly depend on the user's evaluation. What we can say is that, according to the temporal aspects presented in Figure 4, the MST-based playlists, in average, maximize the similarity between subsequent songs, given a specific similarity metric and a musical feature.

Future works include the evaluation of the playlists by systematic tests in the audience. With this initial investigation we hope to define a MIR application that performs queries in a database using as input a song that is not necessary stored in it. Furthermore, the use of shortest-path trees (Dijkstra algorithm) may bring additional insights to the relationship between genres and emotions.

Acknowledgments. Debora C Correa is grateful to FAPESP (2009/50142-0) for financial support, Alexandre L. M. Levada is grateful to CNPq (475054/2011-3) financial support, and Luciano da F. Costa is grateful to CNPq (301303/06-1 and 573583/2008-0) and FAPESP (05/00587-5) for financial support.

References

1. Andric, A., Haus, G.: Automatic playlist generation based on tracking user's listening habits. Multimedia Tools and Applications Journal. 29, Issue 2 (2006)
2. Roads, C. : The Computer Music Tutorial. MIT Press, Massachusetts (1996)
3. Easley. D, Kleinberg, J.: Networks, Crowds and Markets: Reasoning About a Highly Connected World. Cambridge University Press, Cambridge (2010)
4. Miranda, E: Composing with computers. Focal Press, Oxford (2001)
5. Isaacson, D. L., Madsen, R. W.: Markov chains, theory and applications. Krieger Pub Co, Malabar (1976)
6. Clarck, J., Holton, D.A: A First Look at Graph Theory. Word Scientific, Singapore (1991)
7. Alghoniemy, M., Tewfik, A.H.: A network flow model for playlist generation. In: 2001 IEEE International Conf. on Multimedia and Expo, pp. 329–332, Tokyo (2001)
8. Prim, R.C.: Shortest connection networks and some generalizations. Bell System Technical Journal. 36, 1389–1401 (1957)
9. Pauws, S., Eggen, B.: PATS: Realization and User Evaluation of an Automatic Playlist Generator. In: 3rd Int. Simp. on Music Information Retrieval, Paris (2002)
10. Pauws, S., Verhaegh W., Vossen, M.: Music playlist generation by adapted simulated annealing. Information Science. 178, Issue 3, 647–662 (2008)
11. Eerola, T., Toiviainen, P.:MIDI Toolbox: MATLAB Tools for Music Research, University of Jyväskylä (2004)
12. Gabrielsson, A.: The Relationship between Musical Structure and Perceived Expression. In: Hallan, S., Cross, I., Thaut, M. (eds.) The Oxford Handbook of Music Psychology, pp-1041–150. Oxford University Press (2009)
13. Correa, D., Levada, A. L. M, Costa, L. da F.: Finding Community Structure in Music Genres Networks. In: 12th International Society for Music Information Retrieval Conference, pp. 447–452, Miami (2011)

Compression-Based Clustering of Chromagram Data: New Method and Representations

Teppo E. Ahonen

Department of Computer Science
University of Helsinki
teahonen@cs.helsinki.fi

Abstract. We approach the problem of measuring similarity between chromagrams and present two new quantized representations for the task. The first representation is a sequence of optimal transposition index (OTI) values between the global chroma vector and each frame of the chromagram, whereas the second representation uses in similar fashion the global chroma of the query and frames of the target chromagram, thus emphasizing the mutual information of the chromagrams in the representation. The similarity between quantized representations is measured using normalized compression distance (NCD) as the similarity metric, and we experiment with a variant of k-medians algorithm, where the commonly used Euclidean distance has been replaced with NCD, to cluster the chromagrams. The representations and clustering method are evaluated by experimenting how well different cover versions of a composition can be clustered, and based on the experiments, we analyze various parameter settings for the representations. The results are promising and provide possible directions for future work.

Keywords: chromagram, normalized compression distance, clustering

1 Introduction

The chromagram, extracted from the audio data, is a sequence of 12-dimensional vectors that describe the relative energy of the 12 pitch classes of the western tonal scale. The chromagram is robust towards changes in features such as instrumentation and articulation, making it a very suitable feature for various tonality-based similarity measuring tasks. Because of this, chromagram is one of the most commonly used audio features in music information retrieval (MIR); different applications in tasks such as audio classification (e.g. [1]), audio fingerprinting (e.g. [2]), and chord sequence estimation (e.g. [3]) are based on information contained in the chromagram.

Measuring similarity between chromagram representations is essential in various retrieval and classification tasks. Different methods for chromagram similarity measurement have been presented, mostly in the task of cover song identification, where the goal is to determine whether two pieces of music are different renditions of the same composition. Far from trivial, but highly applicable when

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

successful, the task of cover song identification has gained a fair amount of interest from the researchers in the MIR community, yielding a plethora of different representations and similarity metrics. See [4] for an overview of several state-of-the-art methods.

In recent years, a similarity metric called normalized compression distance (NCD) [5] has been succesfully used for parameter-free similarity measuring in various tasks and domains. We apply NCD here, and in order to use the compression-based similarity metric for chromagram data, the continuous chromagram sequences need to be quantized. Several methods of producing a quantized representation from a chromagram exist. The method we use for discretization has been applied in [6], but unlike their work, we are not interested in binary similarity, but instead use the method to produce a sequence of 12 characters that represents the changes in the chromagram during the piece of music.

In this paper we apply the previously discussed discretization method and compression-based similarity metric for the task of clustering chromagram data. In the clustering phase we experiment with a slightly modified variation of the k-medians algorithm, using NCD as the distance metric. We evaluate the performance with real world audio data of cover versions, and examine the effect of smoothing the data with median filtering. The discretization is described in Section 2, and the clustering method is presented in Section 3. The experiments and their results are presented in Section 4, and Section 5 concludes the paper and discusses possible areas of future work.

2 Chroma Contour Representations

In [6], a method for producing a binary similarity matrix between two chromagrams was presented. The method uses optimal transposition index (OTI). OTI calculates the most likely semitone transposition between two chromagrams by first calculating the global chromagrams (i.e. the chromagram frames summed and normalized) and then taking the dot products between one global chromagram and all 12 transpositions of the other. The transposition with the highest dot product value is then used as the most likely semitone transposition between the pieces. Formally, for global chroma vectors G_a and G_b, the OTI is

$$OTI(G_a, G_b) = \arg\max_{0 \leq i \leq M} \{G_a \cdot circshift(G_b, i-1)\}, \quad (1)$$

where M is the maximum of possible transpositions; in our work, this is 12.

In [6], the binary similarity matrix between two pieces of music is obtained by calculating the OTI values between each pair of frames of the two chromagrams, and setting the binary value to the similarity matrix according to the obtained dot product value. We apply the idea, but instead of binary values, we consider all 12 possible transpositions between chroma vectors, and instead of comparing two sequences and constructing a similarity matrix, we transform a continuous chromagram into character sequence representation. We start by calculating the OTI values between the global chroma vector of the piece and each chroma frame

of the piece. Then, the frames are labeled according to the 12 possible OTI values, thus producing a sequence of character labels from an alphabet of size 12. For the lack of a better term, we call this *chroma contour*, as it describes the relative changes between subsequent chromagram vectors. Formally, for a chromagram g_a of length n, and its global chroma vector G_a, the resulting chroma contour sequence (ccs) is

$$ccs(i) = OTI(g_a(i), G_a), \tag{2}$$

where $1 \leq i \leq n$. Finally, each of the 12 possible OTI values in *ccs* are assigned a related symbol.

The representation has the advantage of being completely key-independent, as the sequences produced by OTI are similar in all transpositions. An illustration of a sequence produced by this method is depicted in Figure 1. The similarity is then measured between two such chroma contour sequences. To calculate the chromagrams, we use a window length of 0.1858 seconds with a hop factor of 0.875.

However, this representation only provides information on how the chromatic features change in a single piece of music. This information is clearly useful, but we also wish to consider the relation between two pieces of music, as this is likely beneficial information considering the similarity measuring. The previously presented method can be straightforwardly applied for two pieces, simply by using the global chroma of one piece (the query) and the chroma vectors of the other (the target), producing a similar chroma contour sequence. Again, for the lack of a better term, we call this *cross-chroma contour*. Formally, for a target chromagram g_a of length n and a global query chroma vector G_b, the cross-chroma contour sequence (cccs) is

$$cccs(i) = OTI(g_a(i), G_b), \tag{3}$$

where $1 \leq i \leq n$. Again, finally each of the 12 possible OTI values in *cccs* are assigned a related symbol.

The reasoning for this representation is the idea that if two chromagrams contain similar features, their contours should be highly similar regardless of the global features, whereas two unrelated chromagrams should provide a target cross-chroma contour that does not bear resemblance to the query chroma contour. An illustration of cross-chroma contours for two pieces of music using the global chroma data of the piece of Figure 1 is depicted in Figure 2.

The cross-chroma contour representation is not key invariant, as for example using a semitone transposed global chroma would produce a highly unsimilar sequence in comparison to a sequence produced with the untransposed version. When comparing chromagrams, this needs to be addressed, as the chromagrams extracted from pieces in different keys would be deemed unsimilar regardless of the similarities they share. To overcome this, we first calculate the OTI between the two global chromagrams, then transpose the query according to the OTI value, and then produce the cross-chroma contour sequence. Formally, using notation from Equation 3, this is

$$cccs(i) = OTI(g(i), G_b^n), \tag{4}$$

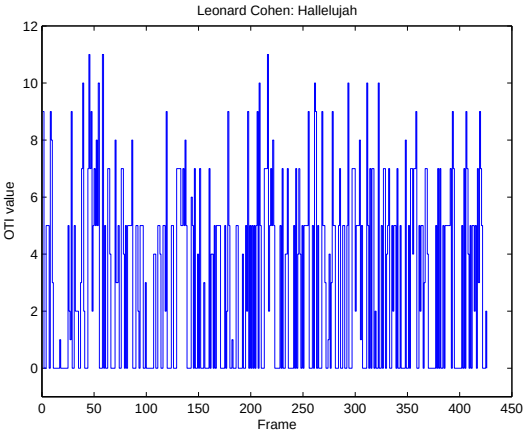


Fig. 1. A chroma contour illustration. For the sake of readability, the sequence depicted here is obtained from a chromagram that was calculated using a quadruple window length.

where G^n is the global chromagram transposed n semitones and $n = OTI(G_a, G_b)$.

3 Clustering Methodology

The clustering method presented in this paper uses normalized compression distance (NCD) [5] as the similarity metric. Using NCD seems appealing for several reasons.

First, NCD is parameter-free, requiring no background information of the data it is applied to. But this, naturally, makes the selection of the data representation crucial for our task: NCD captures the dominant feature similarity between the objects [5] and the representation should therefore express this feature. Second, NCD can be shown to approximate a universal similarity metric, depending on how well the compression algorithm approximates Kolmogorov complexity, the theoretical measure of computational resources needed to produce the object. In addition, NCD has been used succefully for various tasks in music information retrieval (e.g. [7–9]).

To understand what makes NCD a suitable choice as a similarity metric, its background in information theory needs to be explained. Denote $K(x)$ as the Kolmogorov complexity of object x as the length of the smallest program that produces x and $K(x|y)$ as the conditional Kolmogorov complexity, that is, the length of the smallest program that produces x given y as an input. Now, a universal similarity metric called normalized information distance (NID) can be denoted [10]

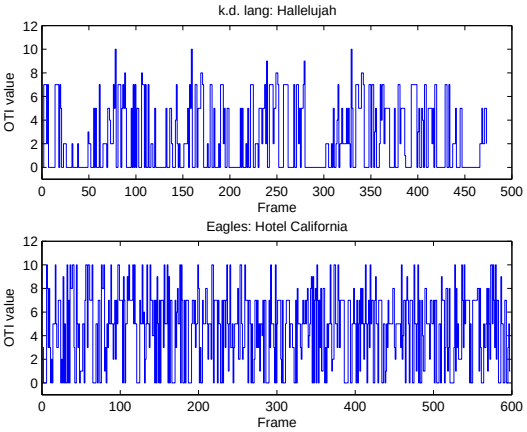


Fig. 2. Two cross-chroma contour illustration, calculated using the global chroma of the piece of music of Figure 1. The upper cross-chroma contour seems to bear more resemblance to the chroma contour of Figure 1. The sequences depicted here are obtained from chromagrams that were calculated using a quadruple window length.

$$NID(x,y) = \frac{\max\{K(x|y), K(y|x)\}}{\max\{K(x), K(y)\}}. \tag{5}$$

NID can be shown to be universal [10], but the incomputability of Kolmogorov complexity makes NID also incomputable. However, the Kolmogorov complexity can be approximated using data compression. This leads to an approximation of NID that is NCD. For two objects x and y, NCD is denoted [5]

$$NCD(x,y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}} \tag{6}$$

where $C(x)$ is the length of x when compressed using a fixed lossless compression algorithm C, and xy is the concatenation of x and y. For the experiments conducted here, we use the bzip2 algorithm for data compression.

3.1 k-medians Clustering

The k-medians clustering method is a variant of the well-known k-means clustering method that uses, as the name implies, median instead of mean when selecting the new cluster centroids. Commonly, Euclidian distance is used as the similarity metric in k-medians clustering, but when clustering symbolized data, the Euclidian distance seems unusable. Here, we apply NCD as the distance measure between the strings and based on the pairwise distances between either chroma or cross-chroma contour representations, conduct the standard k-medians clustering. The strings in the clusters are sorted according to a length-

increasing, lexicographical order, allowing us to select the median from the list of the sorted strings.

4 Evaluation

To evaluate the presented methods and representations, we collected a dataset of cover versions from our personal collections of music. The total dataset consists of 12 ten-song sets of original performances and their cover versions, thus totaling 120 pieces of music. In our experiments, we try to cluster the chromagrams into groups of renditions of the same composition. In practice, this is a cover song identification task expanded with the clustering phase. In order to obtain successful clustering, both the chromagram similarity measuring and the k-medians variant should perform adequately.

To measure the clustering performance we use purity. The purity of a single cluster is calculated as the ratio between the size of the most frequent class (in our case, the cover song set) of the cluster and the size of the whole cluster. The purity for the clustering is then calculated as the average of the single cluster purities. High purity value expresses that the cover versions have been successfully clustered together.

We ran the evaluations for three different-sized datasets. In the *set30* we used three ten-song sets, and in *set60* we doubled the size to six ten-song sets. The *set120* is the complete dataset. For each dataset, the k-medians clustering was run with k of 3, 6, and 12, respectively. The results for the evaluations are presented in Table 1. As the k-medians algorithm selects the initial cluster centroids randomly, we ran the evaluations five times and averaged the results. As a vague baseline comparison, results for a clustering with random distance values is included.

The perfomance of the cross-chroma contour representation is slightly superior to the chroma contour representation. This supports the idea that mutual information between two pieces should be taken into account when producing the representation for a compression-based method. Also, there seems to be robustness in the method, as the performance does not drop significantly as the size of the dataset increases.

4.1 Pre- and Post-Discretization Filtering

The chromagram data extracted from the audio is likely noisy. Often, algorithms that measure chromagram similarity filter the data before applying the similarity measuring. One of the most straightforward ways to smoothen the data is to use median filtering. We experimented with several orders of the median filtering, but here present only the best results, which were obtained with a median filtering of order five.

Also, the sequences produced by our method do oscillate between different OTI values, making the sequences noisy. To make the sequences smoother and thus more compressable, we experimented with median filtering of various orders

for the sequences, but as with the chromagram filtering, present only the results for the best choice of the filtering order, which was also five for the sequence filtering.

Based on the results in Table 1, it seems that both the noise on the chroma data and in the produced character sequences are beneficial for the compression-based similarity metric, and combining both smoothings provides the worst results that are only slightly above, or even below, the random baseline. This could possibly be an outcome of over-simplifying the sequences and losing all distinguishing information. However, on occasional runs the results with filtered data were better than with the unfiltered versions.

Table 1. Purity values of the clustering experiments, averaged over five runs.

	set30	set60	set120
Chroma contour	0.367	0.283	0.217
Cross-chroma contour	**0.374**	**0.317**	**0.257**
Chroma contour and chroma filtering	0.310	0.231	0.162
Cross-chroma contour and chroma filtering	0.344	0.312	0.228
Chroma contour and sequence filtering	0.331	0.258	0.189
Cross-chroma contour and sequence filtering	0.337	0.294	0.212
Chroma contour and both filtering methods	0.133	0.104	0.081
Cross-chroma contour and both filtering methods	0.192	0.162	0.132
Random baseline	0.233	0.117	0.067

5 Conclusions and Future Work

We have presented a method for clustering chromagram data based on the contour of the chromagrams. Our method produces a string representation from the chroma data based on the optimal transposition index values between the global chroma vector and the single chroma vectors of the piece. We also presented a variation where the pairwise mutual information can be utilized by using the global chroma of the query when producing a representation from the target chroma vectors.

Also, the effect of smoothing both the chromagram and the sequence data was experimented, and according to our results, both smoothing methods affect the results negatively. We clustered the chromagrams using a variant of k-medians clustering with normalized compression distance as the similarity metric and used purity to measure the performance of the clustering.

At its best, the method does provide a reasonable level of performance, but it is also clear that several issues still demand consideration. A more fine-grained chroma contour could possibly provide higher cluster purity, as currently several false positives occur and pieces of music end up in wrong clusters. This is likely due to the possibly over-simplified representation; although the representations

are composed using a rather small alphabet, making them thus presumably suitable for the compression algorithm, the trade-off comes as a loss of distinguishing power. A more fine-grained chroma contour could be obtained by either using a chroma representation of 24 or 36 bins, or producing the character sequences using a more complex method instead of the dot product. We are currently working on this, and in addition, experimenting with our k-medians variation, in order to see if there are other features leading to biased results. We are also aware that this paper has not provided comparison with other represenations, similarity metrics, or clustering methods. A thorough comparison with other methodologies is needed in the future.

Acknowledgments. The work presented here was supported by Helsinki Doctoral Programme in Computer Science – Advanced Computing and Intelligent Systems (Hecse). The author wishes to express gratitude to the anonymous reviewers for their insightful comments on the first version of this paper.

References

1. Casey, M., Slaney, M.: The Importance of Sequences in Musical Similarity. In: Proc. ICASSP'06, pp. 5–8. (2006)
2. Bartsch, M., Wakefield, G.: To Catch a Chorus: Using Chroma-based Representations for Audio Thumbnailing. In: Proc. WASPAA'01, pp. 15–18. (2001)
3. Papadopoulos, H., Peeters, G.: Large-Scale Study of Chord Estimation Algorithms Based on Chroma Representation and HMM In: Proc. CBMI'07, pp. 53–60. (2007)
4. Serrà J., Gómez, E., Herrara P.: Audio Cover Song Identification and Similarity: Background, Approaches, Evaluation, and Beyond. In Advances in Music Information Retrieval, pp. 307–332. Springer Verlab (2010)
5. Cilibrasi, R., Vitányi, P.M.B.: Clustering by Compression. IEEE Trans. Information Theory. 51, 1523–1545 (2005)
6. Serrá, J., Gómez E., Herrera P., Serra X.: Chroma Binary Similarity and Local Alignment Applied to Cover Song Identification. IEEE Trans. Audio, Speech and Language Processing. 16, 1138–1151 (2008)
7. Cilibrasi, R., Vitányi, P., De Wolf, R.: Algorithmic Clustering of Music Based on String Compression. Comput. Music J. 28, 49–67 (2004)
8. Li, M., Sleep, R.: Genre Classification Via an LZ78-based String Kernel. In: Proc. ISMIR'05, pp. 252–259. (2005)
9. Bello, J.P.: Grouping Recorded Music by Structural Similarity. In: Proc. ISMIR'09, pp. 531–536. (2009)
10. Li, M., Chen, X., Li, X, Ma, B., Vitányi, P.M.B.: The Similarity Metric. IEEE Trans. Information Theory. 50. 3250–3264 (2004)

GimmeDaBlues: An Intelligent Jazz/Blues Player And Comping Generator for iOS devices

Rui Dias[1], Telmo Marques[2], George Sioros[1], and Carlos Guedes[1]

[1] INESC-Porto / Porto University, Portugal
ruidias74@gmail.com
gsioros@hotmail.com
cguedes@fe.up.pt
[2] CITAR / Universidade Católica do Porto, Portugal
telmomarques33@gmail.com

Abstract. This paper describes an application for iPhone/iPod Touch/iPad devices that allows anyone to play jazz keyboard and solo instruments along a predefined harmonic progression, using the multi-touch properties of the iOS devices. While the user plays keyboard and/or solo instruments, the application automatically generates the bass and drums parts, responding to the user's activity.
Dynamic mapping of the notes and chords available in the graphical interface provides an intuitive and natural way to play otherwise complex chords and scales, while maintaining a physical playability that will be familiar to experienced keyboard players, and provides an entertaining, yet challenging experience for non-musicians.

Keywords: Automatic Music Generation, Blues, Jazz, iOS, Entertainment.

1 Introduction

1.1 GimmeDaBlues Overview

It is widely accepted that Jazz – and this includes the Blues – is a musical style directly related with improvisation and interaction. But how do Jazz musician interact? How do they improvise? It is beyond the extent of this study to answer to such questions, but they trace a new demand (this time pivotal to our application): Could this be transferred to a human-computer environment in a way that either a competent jazz musician or an inexperienced layman could reach the interaction and improvisation proficiency?

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

Existing applications apparently respond partially to these demands. Apps like ThumbJam [1] have an instrument-like approach, providing an interface for expressive playing of sample-based sounds and allowing and flexible mapping of the selected sound. This approach, however, is passive, in that the app doesn't have any kind of information, intelligence or participation on the resulting musical content.
Knowledge from well-known jazz algorithmic improvisation systems like Genjam [2], and Bob [3], provide the base for the creation of intelligent and informed music systems.
GimmeDaBlues — from now on GdB — was developed to fulfil all the purposes for such question: an application for jazz and blues lovers, whether musicians or non-musicians, having playability and ease of use in mind.

For musicians the application can be entertaining and funny. People interact with the bass and the drums while playing the piano or the Hammond-like organ. Improvising simultaneously with the trumpet will give a sense of leadership and control of the entire combo. For non-musicians this can also be challenging and instructive: although GdB has a jazz theory background that contributes to a clean academic performance, timings and groove are entirely up to the user, so that to sound jazzy, the user has to be familiarized with jazz.

1.2 Multi-touch surfaces musical controllers and iOS

In computing interaction within hardware and software, *multi-touch* means a 'touch sensing surface's (trackpad or touchscreen) ability to recognize the presence of two or more points of contact with the surface [4]'. This plural-point awareness allows multiple fingering functionality – surely important in our application – like the use of both hands, both thumb manipulation on the keyboard region, and the *solo-like* fingering on the soloist instrument region defined by the trumpet image. This multi-touch surface method is nowadays widely used by almost every touch screen phones and notebooks, but what makes the iPhone, the iPod Touch, and the iPad devices unique is their quick response to "swipes, pinches, and finger presses [4]. In rhythmical styles such as jazz and blues this features are absolutely primordial for a musical controller. Once these three surfaces are compatible in terms of iOS the application extents to all of them three.

2 Interface

The GdB interface is divided in two main areas in the center, plus the volume controls for the bass and drums on the left and right edges of the screen (Fig. 1). The main areas on the center are the instrument areas. The upper half is the solo instrument while the lower half is the keyboard (chord) instrument. It is not expected to play the right keys especially on such a messy keyboard. Specific keys are not important but moving from left to right on the keyboard results on chords or notes from low to high according to the zone pressed. The small button on the top left drops a menu with options for *Record*, *Setup*, *Library* and *About* screens.

Fig. 1 Main Window

Recording. The Record button starts recording the current session as the user plays. Both the user's instruments and the automatically generated bass and drums will be recorded. When recording is on, a stop button emerges on the top right corner permitting to stop recording whenever we want. Stopping the recording will show a menu with options to save the session in the library or discard it.

Setup. In the *Setup* menu, the user can choose the *Instruments*, *Style*, *Root Key* and *Tempo*. Each option will show the corresponding selection list.
The currently available instruments for the solo (upper) instrument are *Trumpet*, *Bright Trumpet* and *Scoop Down Trumpet*. For the keyboard (lower) instrument the user can choose a *Piano* or a *Blues Organ* sound.
The *Style* option shows a list of the available song styles. Each song style is a well-known jazz/blues chord structure, like a *Classical* (major) *Blues*, or *Minor Blues*.
The *Root Key* changes the song key, transposing all the events in the current style to the selected key.
The *Tempo* option sets the speed in beats-per-minute values.

Library. In the library the user can see and play the available recorded sessions. Each stored recording is a MIDI type 1 file that can easily be exported to the computer using iTunes File Sharing feature. These files can then be opened in any external sequencer or notation program. The file sharing also allow to import files into the GdB app, so the user can easily manage his own session files in the computer and select the ones to send to the mobile device to, for example, go for a given jam session.

3 Technical Description

3.1 Metronome

The underlying beat source is a metronome running at three times the bpm setting. This accounts for the typical triplet subdivision of the beat in traditional jazz and

blues styles. The clock is then divided in two different pipelines, corresponding to two different timelines, synchronized but slightly "out of phase". This is due to a very common practice in jazz playing, which is the anticipation of the first beat. Very often, the musician will anticipate the next first beat while still in the last beat of the current measure, usually by playing a triplet before the beat. This anticipation is not only rhythmical but also harmonic, if the harmony in the next beat changes.

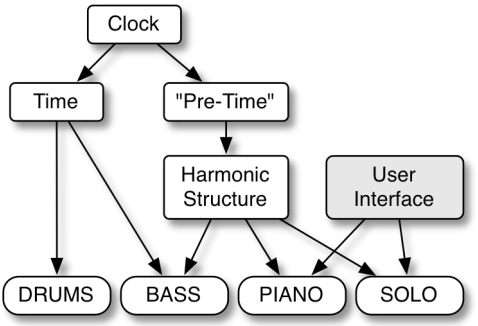


Fig. 2. Control diagram.

The anticipated timeline, the *pre-time*, is used to trigger the harmonic sequence, so if the user plays an anticipated chord up to an eight before the end of the measure, the harmony will correspond to the one in the next measure. The second timeline - the real or *current time*, is the perceived beat and is used to drive the bass and drums algorithms. Technically, this second timeline is a delayed version of the first one. The bpm value can be set in the setup page of the app.

3.2 Harmonic Structure

Each song, called *style* in the app, has a different harmonic structure, drawn from well-known traditional jazz/blues standard chord progressions. The internal sequencer reads the harmonic contents for each measure in real time, triggered by the *pre-time* clock, sending them to the pitched instruments (bass, solo and keyboard). Each instrument then uses the harmonic content information according to their specific algorithm.

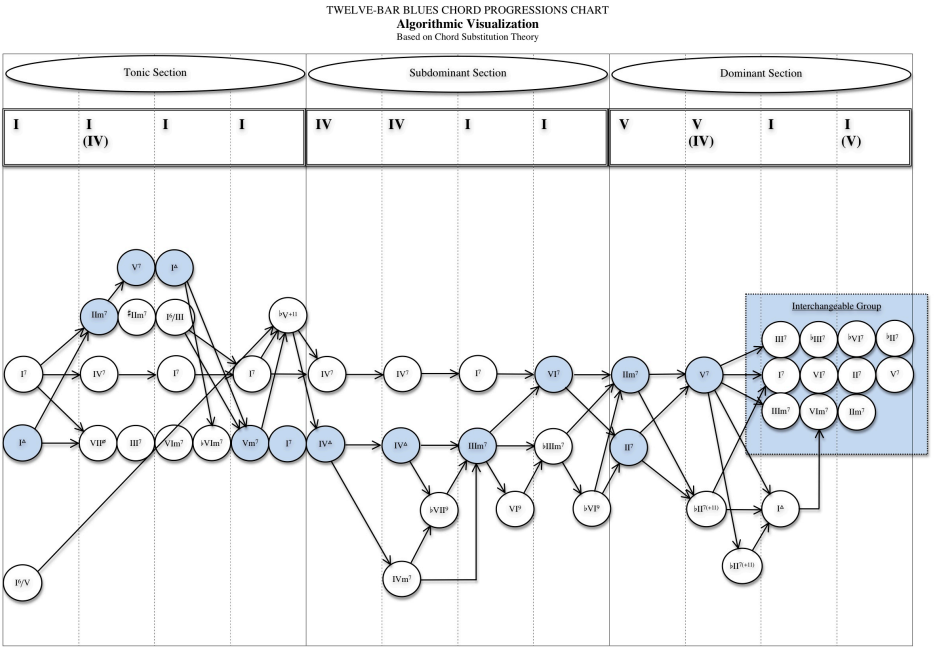


Fig. 3. Interchange chord substitution chart.

The study of the co-relation between common harmonic base progressions in several well-known Blues structures and common interchange and chord substitution procedures was the base of the harmonic structure selections. The chart presented in Fig. 3 was built, developed from jazz theories based on Levine [5], Nettles [6], Steedman [7] & [8], Pease [9] & [10], and Felts [11]. This chart will also be used in the development of an intelligent algorithm for harmonic variation in a future version of the app.

Each *style* defines not only a generic chord progression, but also the *voicing* for each instrument. A *voicing* is the way a given chord is played. In typical jazz and blues performance, the player has complete freedom over the combination of notes he uses to play a chord, as well as their distribution along the instrument's range, using inversions, tensions, and extensions, as long as he maintains a certain coherence with the base chord and/or chord progression. In the GdB app, this notion applies mainly to the keyboard instrument. As for the bass and solo instruments, the *voicing* defined for each song sets a scale that can be played with the corresponding harmony.

3.3 Solo and Keyboard Harmonic Mappings

Each one of the four instruments in GdB has different algorithms, whether relating to the interaction or to the generation method. The pitched instruments use the harmonic contents differently.

Solo instrument. As said before, the solo instrument's voicing data is used to define a scale. Each time a new chord arrives, this scale is mapped dynamically to the horizontal axis of the instrument's corresponding area of the screen.

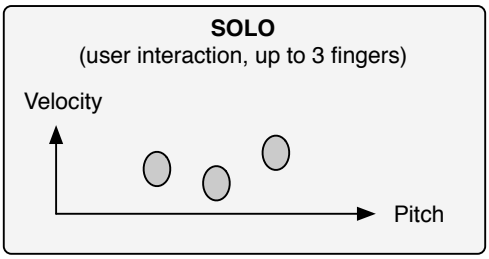


Fig. 4. The solo instrument interface mapping.

Each finger touch in the solo area will produce one single note, corresponding to a given note of the scale associated with the current harmony, that will be sustained while the finger is pressed. Dragging the finger will skim through the scale notes. The vertical position of the finger will determine the attack velocity of the note.

Also, using the multi-touch capabilities of the iOS devices, this area allows for polyphonic events as the user can play with up to three simultaneous fingers. This extra feature is especially interesting for playing short brass section-like riffs in the middle of a solo, or while comping for external soloists.

Keyboard instrument. The keyboard instrument (the lower half of the screen) is built to play block chords. Each finger touch in this area plays a two or three note chord, depending on the position on the vertical axis. By using two fingers, the user can play two chords, like a keyboard player would do, simulating piano playing in a familiar and intuitive way. The combination of the two fingers with two or three notes in each chord allows for the creation of two to six note chords.

The horizontal position in this area also determines the pitch, going from low pitches on the left to high pitches on the right, just like on a piano or organ keyboard, and determines also the chord inversion, so that moving to close positions will have a musically interesting melodic feeling to the chord successions.

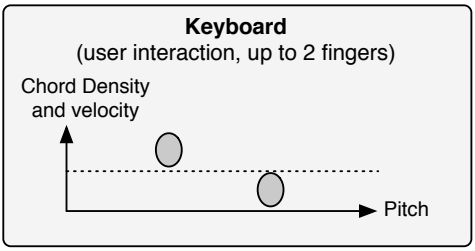


Fig. 5. Keyboard instrument interface mapping

3.4 Bass and Drums

The bass and the drums are generated automatically, according to the harmonic and metrical situation on the twelve bar blues. Once again this is grounded on that model presented in Figure 3**.**
The probabilities change dynamically with the user activity: more events being triggered by the user results in more interaction generated by bass and drums.

Bass. Every time a new chord arrives from the harmonic structure, the chord notes are spread throughout a range of almost three octaves, corresponding loosely to the useful range of a normal bass. These notes form an indexed list of useful notes, which can be played by the bass generator.
The note played in the beat following the chord change will be the chord root note. In the beats where there are no chord changes, the bass will play random notes from the chord.
The note events are triggered according to a probability table, setting each beat of the measure separately. The probabilities change dynamically again based on user activity.

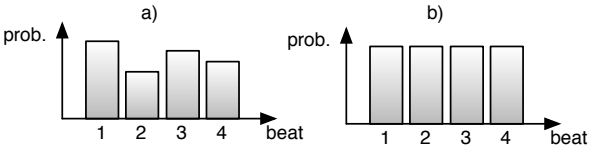


Figure 6. Bass probability table state for each beat at a) minimum user activity; b) maximum user activity

Drums. The drums part is generated automatically by combining a small number of elementary rhythmic patterns, typical for the Blues and Jazz styles. The elementary patterns are played back one at a time but their order is stochastically generated. Each time a pattern finishes playback a new one is selected. The patterns are represented as a binary sequence of triggering or no triggering sound events (see Figure 7). In order to be able to control the density of the resulted rhythm, the elementary patterns are sorted in advance, according to the number of the events each one contains. The probability of a pattern to be selected for playback depends on its position in the ordered list of elementary patterns and on the desired density for the resulted rhythm. The density can vary continuously during performance; however, in practice, only the value at the moment a new pattern is chosen is affecting the resulted rhythm.

Two separate rhythmic patterns are generated, one for the *ride* section, which includes the hi-hat cymbal, the ride cymbal and the crash cymbal, and one for the *snare* section, which includes the snare drum and the kick drum. The elementary patterns have a length either of a whole bar, for the *hi-hat* section, or equal to the beat duration, i.e. a quarter note, for the *snare* section. Since the *snare* patterns are

relatively short, it is not allowed for a pattern to be selected for playback twice in a row. The *snare* patterns must always alternate.

In the *snare* section, only one sound is triggered at a time. The triggered sound is decided stochastically according to relative probabilities that are predefined for the various metrical positions. The sound triggered in most metrical positions is the snare drum sound. Nevertheless, for specific metrical positions, there is a finite probability which ranges between 30 to 80% of replacing the snare drum sound with a kick drum sound. In the *ride* section, the ride cymbal and the crash cymbal are triggered according to a similar algorithm to that for the *snare* section, with the ride cymbal being more often triggered and the crash cymbal being triggered only in specific positions in the 12 bar structure. A hi-hat cymbal sound is triggered according to a fixed pattern.

The amplitudes of the triggered sounds are randomly generated according to predefined MIDI velocity ranges. A different velocity range corresponds to each sound at each metrical position.

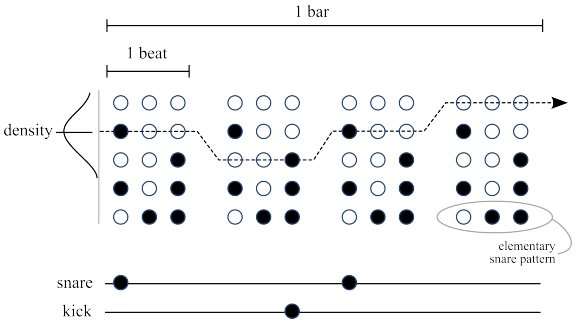


Figure 7. Example of how the *snare* section rhythmic patterns are generated. On every beat a mew elementary pattern is selected for playback according to the density control. Some of the snare drum sounds get replaced by kick drum sounds.

4. Conclusion and Future Developments

Having in mind the use of knowledge from automatic music generation and machine musicianship techniques in the development of an interactive application for music playing led to the development of an iOS application, using the multitouch capabilities to provide a very simple and intuitive, yet powerful, interface.

This interface has to accomplish three fundamental conditions: It must be efficient, user-friendly, and make people want to use it. Dix et al. [12] resumes it in three 'use' words: **Useful**, **Usable**, and **Used**. Several people tested the application throughout two months, musicians and non-musicians, in order to understand which 'use' words had to be improved. We notice that the application is easily usable, but the 'efficiency' and the 'want to use it' aspects could be improved.

Future developments will include the development of an algorithm to produce harmonic variation, according on the afore mentioned theories of chord substitution

rules, the improvement of the walking-bass algorithm and the improvement of the piano voicing mapping algorithm. The possibility of synchronizing several devices wirelessly in order to allow for group playing will also be addressed in a future version.

5. Acknowledgments

This research was done as part of the project "Kinetic controller, driven, adaptive and dynamic music composition systems" funded by the ERDF through the Program COMPETE, by the Portuguese Foundation for Science and Technology (FCT), Project ref. FCOMP-01-0124-FEDER-011414, UTAustin/CD/0052/2008, and partly supported by the European Commission, FP7 (Seventh Framework Programme), ICT-2011.1.5 Networked Media and Search Systems, grant agreement No 287711 (MIReS).

6. References

1. http://thumbjam.com (accessed on June 4th 2011)
2. Biles, J. (1994) "Genjam: A genetic algorithm for generating jazz solos". In *Proceedings of the 1997 ICMC*.
3. Thom, B. (2000) "BoB: an Interactive Improvisational Music Companion" in *Fourth International Conference on Autonomous Agents (Agents-2000)* Barcelona, Spain.
4. Brandon, John (2004) "How the iPhone works" http://www.computerworld.com/s/article/9138644/How_the_iPhone_works (accessed on June 10th 2011)
5. Levine, Mark (1989) *The Jazz Piano Book,* Petaluma: Sher Music
6. Nettles, Barrie & Graf, Richard (2002) *The Chord Scale Theory and Jazz Harmony.* Rottenburg: Advance Music
7. Steedman, Mark (1984) "A Generative Grammar for Jazz Chord Sequences, *Music Perception* 2 (1), pp. 52-77
8. Steedman, Mark (1996) "The Blues and the Abstract Truth: Music and Mental Models" in A. Garnham & J. Oakhill (eds.), *Mental Models In Cognitive Science*. Mahwah, NJ: Erlbaum 1996, 305-318
9. Pease, Ted & Pullig, Ken (2001) Modern Jazz Voicings: Arranging for Small and Medium Ensembles Boston: Berklee Press
10. Pease, Ted (2003) *Jazz Composition: Theory and Practice.* Boston: Berklee Press
11. Felts, Randy (2002) *Reharmonization Techniques.* Boston: Berklee Press
12. Dix, Alan et al. (2004) *Human-Computer Interaction* (3rd. ed). Essex: Pearson and Prentice Hall, p. 5

Oral session 4:
Music Emotion Recognition

Multidisciplinary Perspectives on Music Emotion Recognition: Implications for Content and Context-Based Models

Mathieu Barthet, György Fazekas, and Mark Sandler

Centre for Digital Music
Queen Mary University of London
{mathieu.barthet,gyorgy.fazekas,mark.sandler}@eecs.qmul.ac.uk

Abstract. The prominent status of music in human culture and every day life is due in large part to its striking ability to elicit emotions, which may manifest from slight variation in mood to changes in our physical condition and actions. In this paper, we first review state of the art studies on music and emotions from different disciplines including psychology, musicology and music information retrieval. Based on these studies, we then propose new insights to enhance automated music emotion recognition models.

Keywords: music emotion recognition, mood, metadata, appraisal model

1 Introduction

Since the first empirical works on the relationships between music and emotions [20] [37], a large body of research studies has given strong evidence towards the fact that music can either (i) elicit/induce/evoke emotions in listeners (*felt* emotions), or (ii) express/suggest emotions to listeners (*perceived* emotions), depending on the context [56]. As pointed out by Krumhansl [26], the distinction between *felt* and *perceived* emotions is important both from the theoretical and methodological point of views since the underlying models of representations may differ [71]. One may argue about the fact that music can communicate and trigger emotions in listeners and this has been the subject of numerous debates [37]. However a straightforward demonstration of the latter doe not require a controlled laboratory setting and may be conducted in a common situation, at least in certain cultures, that of watching/listening movies with accompanying soundtracks. In the documentary on film score composer Bernard Hermann [61], the motion picture editor Paul Hirsch (e.g. Star Wars, Carrie) discusses the effect of music in a scene from Alfred Hitchcock's well-known thriller/horror movie *Psycho*, whose soundtrack was composed by Hermann: *"The scene consisted of three very simple shots, there was a close up of her [Janet Lee] driving, there was a point of view of the road in front of her and there was a point of view of the police car behind her that was reflected in the rear mirror. The material was so*

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

simple and yet the scene was absolutely gripping. And I reached over and I turned off the sound to the television set and I realised that the extreme emotional duress I was experiencing was due almost entirely to the music.". With regard to music retrieval, several studies on music information needs and user behaviors have stimulated interest in developing models for the automatic classification of music pieces according to the emotions or mood they suggest. In [28], the responses of 427 participants to the question *"When you search for music or music information, how likely are you to use the following search/browse options?"* showed that emotional/mood states would be used in every third song query, should they be possible. The importance of musical mood metadata was further confirmed in the investigations by Lesaffre et al. [30] which give high importance to affective/emotive descriptors, and indicate that users enjoy discovering new music by entering mood-based queries, as well as those by Bischoff et al. [5] which showed that 15% of the song queries on the web music service Last.fm were made using mood tags. As part of our project Making Musical Mood Metadata (M4) in partnership with the BBC and I Like Music, the present study aims to (i) review the current trends in music emotion recognition (MER), and (ii) provide insights to improve MER models. The remainder of this article is organised as follows. In Section 2, we present the three main types of (music) emotion representations (categorical, dimensional and appraisal). In Section 3, we review MER studies by focusing on those published between 2009 and 2011, and discuss the current trends in terms of features and feature selection frameworks. Section 4 presents state-of-the-art's machine learning techniques for MER. In Section 5, we discuss some of the findings in MER and conclude by highlighting the main implications to improve content and context-based MER models.

2 Representation of Emotions

2.1 Categorical Model

Table 1 presents the main categorical and dimensional emotion models used in the MER studies reviewed in this article. According to the categorial approach, emotions can be represented as a set of categories that are distinct from each others. Ekman's categorical emotion theory [13] introduced *basic* or universal emotions that are expected to have prototypical facial expressions and emotion-specific physiological signatures. The seminal work from Hevner [21] highlighted (i) the bipolar nature of music emotions (e.g. happy/sad), (ii) a possible way of representing them spatially across a circle, as well as (iii) the multi-class and multi-label nature of music emotion classification. Schubert proposed a new taxonomy, the updated Hevner model (UHM) [54], which refined the set of adjectives proposed by Hevner, based on a survey conducted by 133 musically experienced participants. Based on Hevner's list, Russell's circumplex of emotion [44], and Whissell's dictionary of affect [65], the UHM consists in 46 words grouped into nine clusters.

Bischoff et al. [6] and Wang et al. [63] proposed categorical emotion models by dividing the Thayer-Russell Arousal/Valence space (see Section 2.2) into into

Table 1. Categorical and dimensional models of music emotions used in MER. Cat.: Categorical; Dim.: Dimensional; Ref.: References.

Notation	Description	Approach	Ref.
UHM9	Update of Hevner's adjective Model (UHM) including nine categories	Cat.	[54]
AMC5C	5 MIREX audio mood classification (AMC) clusters ("Passionate","Rollicking", "Literate", "Humorous", "Aggressive")	Cat.	[22] [9] [6] [58] [62]
5BE	5 basic emotions ("Happy", "Sad", "Tender", "Scary", "Angry")	Cat.	[12] [45]
AV4Q	4 quadrants of the Thayer-Russell AV space ("Exuberance", "Anxious/Frantic", "Depression", "Contentment")	Cat.	[6] [63]
AV11C	11 subdivisions of the Thayer-Russell AV space ("Pleased", "Happy", "Excited", "Angry", "Nervous", "Bored", "Sad", "Sleepy", "Peaceful", "Relaxed", and "Calm")	Cat.	[19]
AMG12C	12 clusters based on AMG tags	Cat.	[33]
72TCAL500	72 tags from the CAL-500 dataset (genres, instruments, emotions, etc.)	Cat.	[4]
AV4Q-UHM9	Categorisation of UHM9 in Thayer-Russell's quadrants (AV4Q)	Cat.	[40]
AV8C	8 subdivisions of the Thayer-Russell AV space	Cat.	[24]
4BE	4 basic emotions ("Happy", "Sad", "Angry", "Fearful")	Cat.	[59]
4BE-AV	4 basic emotions based on the AV space ("Happy", "Sad", "Angry", "Relaxing")	Cat.	[63]
9AD	Nine affective dimensions from Asmus ("Evil", "Sensual", "Potency", "Humor", "Pastoral", "Longing", "Depression", "Sedative", and "Activity")	Dim.	[2]
AV	Arousal/Valence (Thayer-Russell model)	Dim.	[19]
EPA	Evaluation, potency, and activity (Osgood model)	Dim.	
6D-EPA	6 dim. correlated with the EPA model	Dim.	[35]
AVT	Arousal, valence, and tension	Dim.	[12]

four quadrants (AV4Q). [19] proposed subdivisions of the four AV space quadrants into a larger set, composed of 11 categories (AV11C). Their model, assessed on a prototypical database, led to high MER performance (see Section 3). [22] and [33] proposed mood taxonomies based on the (semi-)automatic analysis of mood tags with clustering techniques. [22] applied an agglomerative hierarchical clustering procedure (Ward's criterion) on similarity data between mood labels mined from the `AllMusicGuide.com` (AMG) website presenting annotations made by professional editors. The procedure generated a set of five clusters which further served as a mood representation model (denoted AMC5C, here) in the MIREX audio mood classification task and has been widely used since (e.g. in [22], [9], [6], and [62]). In this model, the similarity between emotion labels is computed from the frequency of their co-occurence in the dataset. Consequently some of the mood tag clusters may comprise tags which suggest different emotions. Training MER models on these clusters may be misleading for inference systems, as shown in [6] where prominent confusion patterns between clusters are reported (between Clusters 1 and 2, as well as between Clusters 4 and 3). [24] proposed a new categorical model by collecting 4460 mood tags and AV values from 10 music clip annotators and by further grouping them relying on unsupervised classification techniques. The collected mood tags were processed to get rid of synonymous and ambiguous terms. Based on the frequency distribution of the 115 remaining mood tags, the 32 most frequently used tags were retained. The AV values associated with the tags were processed using K-means clustering which led to a configuration of eight clusters (AV8C). The results show that some regions can be identified by the same representative mood tags

as in previous models, but that some of the mood tags present overlap between regions. Categorical approaches have been criticized for their restrictions due to the discretization of the problem into a set of "families" or "landmarks" [39] [8], which prevent to consider emotions which differ from these landmarks. However, as highlighted in the introduction, for music retrieval applications based on language queries, such landmarks (keywords/tags) have shown to be useful.

2.2 Dimensional Model

In contrast to categorical emotion models, dimensional models characterise emotions based on a small number of dimensions intended to correspond to the internal human representation of emotions. The psychologist Osgood [41] devised a technique for measuring the connotative meaning of concepts, called the *semantic differential technique* (SDT). Experiments were conducted with 200 undergraduate students who were asked to rate 20 concepts using 50 descriptive scales (7-point Likert scales whose poles were bipolar adjectives) [41]. Factor analyses accounted for almost 70% of the common variance in a three-dimensional configuration (50% of the total variance remained unexplained). The first factor was clearly identifiable as *evaluative*, for instance representing adjective pairs such as *good/bad*, *beautiful/ugly* (dimension also called *valence*), the second factor identified fairly well as *potency*, for instance related to bipolar adjectives *large/small*, *strong/weak*, *heavy/light* (dimension also called *dominance*), and the third factor appeared to be mainly an *activity* variable, related to adjectives such as *fast/slow*, *active/passive*, *hot/cold* (dimension also called *arousal*). Osgood's EPA model was used for instance in the study [10] investigating how well music (theme tune) can aid automatic classification of TV programmes from BBC Information & Archive. A slight variation of the EPA model was used in [11] with the *potency* dimension being replaced by one related to *tension*. Although Osgood's model has been shown to be relevant to classify affective concepts, its adaptability to music emotions is notwithstanding not straightforward. Asmus [2] replicated Osgood's SDT in the context of music emotions classification. Measures were developed from 2057 participants on 99 affect terms in response to musical excerpts and then factor analysed. Nine affective dimensions (9AD) were found to best represent the measures, two of which were found to be common to the EPA model. Probably because it is harder to visually represent nine dimensions and because it complicates the classification problem, this model has not been used yet in the MIR domain, to our knowledge.

The works that have had the most influence on the choice of emotion representations in MER so far are those from Russell [44] and Thayer [57]. Russell devised a *circumplex model of affect* which consists of a two-dimensional, circular structure involving the dimensions of *arousal* and *valence* (denoted AV and called the *core affect dimensions* following Russell's terminology). Within the AV model, emotions that are across a circle from one another correlate inversely, aspect which is also in line with the semantic differential approach and the bipolar adjectives proposed by Osgood. Schubert [53] developed a measurement interface called the "two-dimensional emotional space" (2DES) using Russell's core affect dimensions and proved the validity of the methodology, experimentally. While

the AV space stood out amongst other models for its simplicity and robustness, higher dimensionality have shown to be needed when seeking for completeness. The potency or dominance dimension related to power and control proposed by Osgood is necessary to make important distinctions between fear and anger, for instance, which are both active and negative states. Fontaine et al. [16] advocated the use of a fourth dimension related to the expectedness or unexpectedness of events, which to our knowledge has not been applied in the MIR domain so far.

A comparison between the categorical, or discrete, and dimensional models has been conducted in [11]. Linear mapping techniques revealed a high correspondence along the core affect dimensions (arousal and valence), and the three obtained dimensions could be reduced to two without significantly reducing the goodness of fit. The major difference between the discrete and categorical models concerned the poorer resolution of the discrete model in characterizing emotionally ambiguous examples. [60] compared the applicability of music-specific and general emotion models, the Geneva Emotional Music Scale (GEMS) [71], the discrete and dimensional AV emotion models, in the assessment of music-induced emotions. The AV model outperformed the other two models in the discrimination of music excerpts, and principal component analysis revealed that 89.9% of the variance in the mean ratings of all the scales (in all three models) was accounted for by two principal components that could be labelled as valence and arousal. The results also revealed that personality-related differences were the most pronounced in the case of the discrete emotion model, aspect which seems to contradict that obtained in [11].

2.3 Appraisal Model

The appraisal approach was first advocated by Arnold [1] who defined appraisal as a cognitive evaluation able to distinguish qualitatively among different emotions. The theory of appraisal therefore accounts for individual differences and variations to responses across time [43], as well as cultural differences [47]. The component process appraisal model (CPM) [48] describes an emotion as a process involving five functional components: cognitive, peripheral efference, motivational, motor expression, and subjective feeling. Banse and Scherer [3] proved the relevance of CPM predictions based on acoustical features of vocal expressions of emotions. Significant correlations between appraisals and acoustic features were also reported in [27] showing that inferred appraisals were in line with the theoretical predictions. Mortillaro et al. [39] advocate that the appraisal framework would help to address the following concerns in automatic emotion recognition: (i) how to establish a link between models of emotion recognition and emotion production? (ii) how to add contextual information to systems of emotion recognition? (iii) how to increase the sensitivity with which weak, subtle, or complex emotion states can be detected? All these points are highly significant for MER with a MIR perspective whereas appraisal models such as the CPM have not yet been applied in the MIR field, to our knowledge. The appraisal framework is especially promising for the development of context-sensitive automatic emotion recognition systems taking into account the environment (e.g. work, or home), the situation (relaxing, performing a task), or the subject (personnality traits),

for instance [39]. This comes from the fact that appraisals themselves represent abstractions of contextual information. By inferring appraisals (e.g. obstruction) from behaviors (e.g. frowning), information about causes of emotions (e.g. anger) can be inferred [7].

3 Acoustical and Contextual Analysis of Emotions

Studies in music psychology [56], musicology [18] and music information retrieval [25] have shown that music emotions were related to different musical variables. Table 2 lists the content and context-based features used in the studies reviewed hereby. Various acoustical correlates of articulation, dynamics, harmony, instrumentation, key, mode, pitch, melody, register, rhythm, tempo, musical structure, and timbre have been used in MER models. Timbre features have shown to provide the best performance in MER systems when used as individual features [52] [73]. Schmidt et al. investigated the use of multiple audio content-based features (timbre and chroma domains) both individually and in combination in a feature fusion system [52] [49]. The best individual features were octave-based spectral contrast and MFCCs. However, the best overall results were achieved using a combination of features, as in [73] (combination of rhythm, timbre and pitch features). Eerola et al. [12] extracted features representing six different musical variables (dynamics, timbre, harmony, register, rhythm, and articulation) to further apply statistical feature selection (FS) methods: multiple linear regression (MLR) with a stepwise FS principle, principle component analysis (PCA) followed by the selection of an optimal number of components, and partial least square regression (PLSR) with a Bayesian information criterion (BIC) to select the optimal number of features. PLSR simultaneously allowed to reduce the data while maximising the covariance between the features and the predicted data, providing the highest prediction rate (R^2=.7) with only two components. However, feature selection frameworks operating by considering all the emotion categories or dimensions at the same time may not be optimal; for instance, features explaining why a song expresses "anger" or why another sounds "innocent" may not be the same. Pairwise classification strategies have been successfully applied to musical instrument recognition [14] showing the interest of adapting the feature sets to discriminate two specific instruments. It would be worth investigating if music emotion recognition could benefit from pairwise feature selection strategies as well.

In addition to audio content features, lyrics have also been used in MER, either individually, or in combination with features belonging to different domains (see multi-modal approaches in Section 4.4). Access to lyrics has been facilitated by the emergence of lyrics databases on the web (e.g. lyricwiki.org, musixmatch.com), some of them providing APIs to retrieve the data. Lyrics can be analysed using standard natural language processing (NLP) techniques. To characterise the importance of a given word in a song given the corpus it belongs to, authors used the term frequency - inverse document frequency (TF-IDF) measure [9] [36]. Methods to analyse emotions in lyrics have been developed using lexical resources for opinion and sentiment mining such as SentiWordNet

Table 2. Content (audio and lyrics) and context-based features used in MER (studies between 2009 and 2011)

Type	Notation	Description	References
		Content-based features	
Articulation	EVENTD	Event density	[12]
Articulation/Timbre	ATTACS	Attack slope	[12]
Articulation/Timbre	ATTACT	Attack time	[12]
Dynamics	AVGENER	Average energy	[19]
Dynamics	INT	Intensity	[40]
Dynamics	INTR	Intensity ratio	[40]
Dynamics	DYN	Dynamics features	[45]
Dynamics	RMS	Root mean square energy	[12] [35] [45]
Dynamics	LOWENER	Low energy	[35]
Dynamics	ENER	Energy features	[36]
Harmony	OSPECENT	Octave spectrum entropy	[12]
Harmony	HARMC	Harmonic change	[12]
Harmony	CHROM	Chroma features	[52]
Harmony	HARMF	Harmony features	[45]
Harmony	RCHORDF	Relative chord frequency	[55]
Harmony	WCHORDD	Weighted chord differential	[35]
Instrum./Rhythm	PERCTO	Percussion template occurrence	[58]
Instrumentation	BASSTD	Bass-line template distance	[58]
Key/Mode	KEY	Key	[19]
Key/Mode	KEYC	Key clarity	[12]
Key/Mode	MAJ	Majorness	[12]
Key/Mode	SPITCH	Salient pitch	[12]
Key/Mode	WTON	Weighted tonality	[35]
Key/Mode	WTOND	Weighted tonality differential	[35]
Pitch/Melody	PITCHMIDI	Pitch MIDI features	[73]
Pitch/Melody	MELOMIDI	Melody MIDI features	[73]
Pitch/Melody	PITCH	Pitch features	[45]
Pitch/Timbre	ZCR	Zero-crossing rate	[73] [72]
Register	CHROMD	Chromagram deviation	[12]
Register	CHROMC	Chromagram centroid	[12]
Rhythm/Tempo	BEATINT	Beat interval	[19]
Rhythm/Tempo	SPECFLUCT	Spectrum fluctuation	[12]
Rhythm/Tempo	TEMP	Tempo	[12]
Rhythm/Tempo	PULSC	Pulse clarity	[12]
Rhythm/Tempo	RHYCONT	Rhythm content features	[73]
Rhythm/Tempo	RHYSTR	Rhythm strength	[40]
Rhythm/Tempo	CORRPEA	Correlation peak	[40]
Rhythm/Tempo	ONSF	Onset frequency	[40]
Rhythm/Tempo	RHYT	Rhythm features	[45]
Rhythm/Tempo	SCHERHYT	Scheirer rhythm features	[55]
Rhythm/Tempo	PERCF	Percussive features	[36]
Structure	MSTRUCT	Multidimensional structure features	[12]
Structure	STRUCT	Structure features	[45]
Timbre	SPECC	Spectral centroid	[6] [35] [73]
Timbre	HARMSTR	Harmonic strength	[19]
Timbre	MFCC	Mel frequency cepstral coefficient	[6] [4] [58] [73] [62] [45] [52] [49] [72] [59] [51] [45]
Timbre	SPECC	Spectral centroid	[12] [73] [72] [50] [52] [40] [55]
Timbre	SPECS	Spectral spread	[12]
Timbre	SPECENT	Spectral entropy	[12]
Timbre	SPECR	Spectral rolloff	[12] [73] [72] [50] [52] [40] [55]
Timbre	SF	Spectral flux	[73] [72] [50] [52] [40] [55]
Timbre	OBSC	Octave-based spectral contrast	[50] [52] [49] [51] [40] [29]
Timbre	RPEAKVAL	Ratio between average peak and valley strength	[40]
Timbre	ROUG	Roughness	[12]
Timbre	TIM	Timbre features	[45]
Timbre	SPEC	Spectral features	[36]
Timbre	ECNTT	Echo Nest timbre feature	[51] [36]
Lyrics	SENTIWORD	Occurence of sentiment word	[9]
Lyrics	NEG-SENTIW	Occurrence of sentiment word with negation	[9]
Lyrics	MOD-SENTIW	Occurrence of sentiment word with modifier	[9]
Lyrics	WORDW	Word weight	[9]
Lyrics	LYRIC	Lyrics feature	[73]
Lyrics	RSTEMFR	Relative stem frequency	[55]
Lyrics	TF-IDF	Term frequency - Inverse document frequency	[9] [36]
Lyrics	RHYME	Rhyme feature	[63]
		Context-based features	
Social tags	TAGS	Tag relevance score	[4]
Web-mined tags	DOCRS	Document relevance score	[4]
Metadata	ARTISTW	Artist weight	[9]
Metadata	META	Metadata features (e.g. artist's name, title)	[55]

(measures of positivity, negativity, objectivity) [9], and the affective norm for English words (measures of arousal, valence, and dominance) [36]. Since meaning emerges from subtle word combinations and sentence structure, research is still needed to develop new features characterising emotional meanings in lyrics. [63] proposed a feature to characterise rhymes whose patterns are relevant to emotion expression, as poems can attest. To attempt to improve the performance of MER systems only relying on content-based features, and in order to bridge the semantic gap between the raw data (signals) and high-level semantics (meanings), several studies introduced context-based features. [9], [6], [4], and [62] used music tags mined from websites known to have good quality information about songs, albums or artists (e.g. bbc.co.uk, rollingstone.com), social music platform (e.g. last.fm), or web blogs (e.g. livejournal.com). Social tags are generally fused with audio features to improve overall performance of the classication task [6] [4] [62].

4 Machine Learning for Music Emotion Recognition

4.1 Early Categorical Approaches and Multi-Label Classification

Associating music with discrete emotion categories was demonstrated by the first works that used an audio-based approach. Li et al. [31] used a song database hand-labelled with adjectives belonging to one of 13 categories and trained Support Vector Machines (SVM) on timbral, rhythmic and pitch features. The authors report large variation in the accuracy of estimating the different mood categories, with the overall accuracy (F score) remaining below 50%. Feng et al. [15] used a Back Propagation Neural Network (BPNN) to recognise to which extent music pieces belong to four emotion categories ("happiness", "sadness", "anger", and "fear"). They used features related to tempo (fast-slow) and articulation (staccato-legato), and report 66% and 67% precision and recall, respectively. However, the actual accuracy of detecting each emotion fluctuated considerably. The modest results obtained with early categorical approaches can be attributed to the difficulty in assigning music pieces to any single category, and the ambiguity of mood adjectives themselves. For these reasons subsequent research have moved on to use multi-label, fuzzy or continuous (dimensional) emotion models.

In multi-label classification, training examples are assigned multiple labels from a set of disjoint categories. MER was first formulated as a multi-label classification problem by Wieczorkowska et al. [66] applying a classifier specifically adopted to this task. In a recent study, Sanden and Zhang [46] examined multi-label classification in the general music tagging context (emotion labelling is seen as a subset of this task). Two datasets, the CAL500 and approximately 21,000 clips from Magnatune (each associated with one or more of 188 different tags) were used in the experiments. The clips were modeled using statistical distributions of spectral, timbral and beat features. The authors tested Multi-Label k-Nearest Neighbours (MLkNN), Calibrated Label Ranking (CLR), Backpropagation for Multi-Label Learning (BPMLL), Hierarchy of Multi-Label Classifiers (HOMER), Instance Based Logistic Regression (IBLR) and Binary Relevance

Table 3. Content-based music emotion recognition (MER) models (studies between 2009 and 2011). [a]: *F-measure;* [b]: *Accuracy;* [c]: *Coefficient of determination* R^2*;*[d]*: Average Kullback-Leibler divergence;* [e]*: Average distance;* [f]*: Mean* l^2 *error.* SSD: statistical spectrum descriptors. BAYN: Bayesian network. ACORR: Autocorrelation. Best reported configurations are indicated in bold.

Reference	Modalities	Dtb (# songs)	Model (notation)	Decision hor.	Features (no.)	Machine learn.	Perf.
Lin et al. (2009) [33]	Audio	AMG (1535)	Cat. (AMG12C)	track	MARSYAS (436)	SVM	56.00%[a]
Han et al. (2009) [19]	Audio	AMG (165)	Cat. (AV11C)	track	KEY, AVGENER, TEMP, σ(BEATINT), σ(HARMSTR)	**SVR**, SVM, GMM	94.55%[b]
Eerola et al. (2009) [12]	Audio	Soundtrack110 (110)	Cat. (5BE) & Dim. (AV & AVT)	15.3 s (avg)	RMS, SPECC, SPECS, SPECENT, ROUG, OSPECENT, HARMC, KEYC, MAJ, CHROMC, CHROMD, SPITCH, SPECFLUCT, TEMP, PULSC, EVENTD, ATTACS, ATTACT, MSTRUCT (29)	MLR + STEPS, PCA + FS, **PLSR + DT**	70%[c] (avg)
Tsunoo et al. (2010) [58]	Audio	CAL500 (240)	Cat. (AMC5C)	track	PERCTO (4), BASSTD (80), 26 M,σ MFCCs, 12 M,σ corr(Chroma)	**TEML + SVM**	56.4%[d]
Zhao et al. (2010) [73]	Audio	Chin. & West. (24)	Cat. (AV4Q)	30s	**PITCH (5), RHYT (6), MFCCs (10), SSDs (9)**	**BAYN**	74.9%[b]
Schmidt et al. (2010) [50]	Audio	MoodSwings Lite (240)	Dim. (AV)	1s	OBSC	MLR, LDS Kalman, LDS KALF, **LDS KALFM**	2.88[d]
Schmidt et al. (2010) [52]	Audio	MoodSwings Lite (240)	Cat. (AV4Q) & Dim. (AV)	1s	**MFCCs**, CHROM (12), SSDs, OBSC	SVM / PLSR, **SVR**	0.137[e]
Schmidt & Kim (2010) [49]	Audio	MoodSwings Lite (240)	Dim. (AV)	15s / 1s	**MFCCs**, ACORR(CHROM), SSDs, **OBSC**	MLR, PLSR, **SVR**	3.186 / 13.61[d]
Myint & Pwint (2010) [40]	Audio	Western pop (100)	Cat. (AV4Q-UHM9)	segment	INT, INTR, SSD, OBSC, RHYSTR, CORRPEA, RPEAKVAL, M(TEMP), M(ONSF)	OAO FSVM	37%[b]
Lee et al. (2011) [29]	Audio	Clips (1000)	Dim. 2 (AV)	20s	OBSC	**SVM**	67.5%[b]
Mann et al. (2011) [35]	Audio	TV theme tunes (144)	Dim. (6D-EPA)	track	RMS, LOWENER, SPECC, WTON, WTOND, WCHORDD, TEMP	**SVM**	80-94%[b]
Vaizman et al. (2011) [59]	Audio	Piano, Vocal (76)	Cat. (4BE)	track	34 MFCCs	DTM	60%[a]
Schmidt & Kim (2011) [51]	Audio	MoodSwings Lite (240)	Dim. (AV)	15s / 1s	**MFCCs (20)**, OBSC, ECNTTs (12)	MLR, **CRF**	0.122[f]
Saari et al. (2011) [45]	Audio	Film soundtrack (104)	Cat. (5BE)	track	52 (DYN, RHY, PITCH, HARM, TIM, STRUCT) + MFCCs (14)	**NB**, k-NN, SVM, SMO	59.4%[b]
Wang et al. (2011) [63]	Lyrics	Chinese songs (500)	Cat. (4BE-AV)	track	TF-IDF, RHYME	MLR, NB, SVM-SMO, DECT (J48)	61.5%[a]

Table 4. Multi-modal music emotion recognition (MER) models (studies between 2009 and 2011). [a]*: F-measure;* [b]*: Accuracy;* [c]*: Mean average precision;* [d]*: Coefficient of determination* R^2. FSS: Feature subset selection. Best reported configurations are indicated in bold.

Reference	Modalities	Dtb (# songs)	Model (notation)	Decision hor.	Features (no.)	Machine learn.	Perf.
Dang & Shirai (2009) [9]	Lyrics, Web-mined Tags	LiveJournal, LyricWiki (6000)	Cat. (AMC5C)	track	TF/IDF, SENTIWORD, NEG-SENTIW, MOD-SENTIW, WORDW, ARTISTW	SVM, **NB**, Graph-based	57.44%[b]
Bischoff et al. (2009) [6]	Audio, Social tags	Last.fm, AMG (1192)	Cat. (AMC5C) & AV4Q	30s	MFCCs, TEMP, CHROM (12), SPECC, ... / log(TF)	**SVM (RBF)**, LOGR, RANF, GMM, K-NN, DECT, **NB**	57.2%[a]
Barrington et al. (2009) [4]	Audio, Social tags, Web-mined tags	Last.fm, CAL500 (500)	Cat. (72TCAL500)	30s	MFCCs (39), Δ MFCCs, $\Delta\Delta$ MFCCs, CHROM (12) / + 8-GMM, TAGRS, DOCRS	**CSA**, RANB, KC-SVM	53.8%[c]
Wang et al. (2010) [62]	Audio, Social tags	Last.fm, WordNet, AMG (1804)	Cat. (AMC5C)	track	MARSYAS (138) & PSYSOUND3 + FSS / MFCCs + GMM	SVM PPK-RBF / NRQL	60.6%[b]
Zhao et al. (2010) [73]	Audio, Lyrics, MIDI	Chinese songs (500)	Cat. (AV4Q)	track	MFCCs, LPC, SPECC, SPECR, SPECF, ZCR, ... (113) / N-GRAM LYRIC (2000) / PITCH-MIDI, MELOMIDI (101)	**SVM**, NB, DECT	61.6%[b]
Schuller et al. (2011) [55]	Audio, Lyrics, Metadata	NTWICM, lyricsDB, LyricWiki (2648)	Dim. (AV)	track	RCHORDF (22), SCHERHYT (87), SPECC,... (24) / RSTEMFR (393), META (152)	**ConceptNet, Porter stemming, UREPT**	.60 (A) & .74 (V)[d]
McVicar et al. (2011) [36]	Audio, Lyrics	EchoNest API, lyric-smode.com, ANEW (119 664)	Dim.(AV)	track	TF-IDF, ECNT (65)	**CCA**	N/A

kNN (BRkNN) models, and two separate evaluations were performed using the two datasets. In both cases, the CLR classifier using a Support Vector Machine (CLR_{SVM}) outperformed all other approaches (peak F_1 score of 0.497 and precision of 0.642 on CAL500). However, CLR with Decision Trees, BPMLL, and MLkNN also performed competitively.

4.2 Fuzzy Classification and Emotion Regression

A possible approach to account for subjectivity in emotional responses is the use of fuzzy classification incorporating fuzzy logic into conventional classification strategies. The work of Yang et al. [70] was the first to take this route. As opposed to associating pieces with a single or a discrete set of emotions, fuzzy classification uses fuzzy vectors whose elements represent the likelihood of a piece belonging to each respective emotion categories in a particular model. In [70], two classifiers, Fuzzy k-NN (FkNN) and Fuzzy Nearest Mean (FNM), were tested using a database of 243 popular songs and 15 acoustic features. The authors performed 10-fold cross validation and reported 68.22% and 70.88% mean accuracy for the two classifiers respectively. After applying stepwise backward feature selection, the results improved to 70.88% and 78.33%.

The techniques mentioned so far rely on the idea that emotions may be organised in a simple taxonomy consisting of a small set of universal emotions (e.g. happy or sad) and more subtle differences within these categories. Limitations of this model include *i)* the fixed set of classes considered, *ii)* the ambiguity in the meaning of adjectives associated with emotion categories, and *iii)* the potential heterogeneity in the taxonomical organisation. The use of a continuous emotion space such as Thayer-Russell's Arousal-Valence (AV) space and corresponding dimensional models is a solution to these problems. In the first study that addresses these issues [69], MER was formulated as a regression problem to map high-dimensional features extracted from audio to the two-dimensional AV space directly. AV values for *induced* emotion were collected from 253 subjects for 195 popular recordings. After basic dimensionality reduction of the feature space, three regressors were trained and tested: Multiple Linear Regression (MLR) as baseline, Support Vector Regression (SVR) and Adaboost.RT, a regression tree ensemble. The authors reported coefficient of determination statistics (R^2) with peak performance of 58.3% for arousal, and 28.1% for valence using SVR. Han et al. [19] used SVR for training distinct regressors to predict arousal and valence both in terms of Cartesian and polar coordinates of the AV space. A policy for partitioning the AV space (AV11C) and mapping coordinates to discrete emotions was used, and an increase in accuracy from 63.03% to 94.55% was obtained when polar coordinates were used in this process. Notably Gaussian Mixture Model (GMM) classifiers performed competitively in this study. Schmidt et al. [52] showed that Multi-Level Least-Squares Regression (MLSR) performs comparably to SVR at a lower computational cost. An interesting observation is that combining multiple feature sets does not necessarily improve regressor performance, probably due to the curse of dimensionality. The solution was seen in the use of different fusion topologies, i.e. using separate regressors for each

feature set. Huq et al. [23] performed a systematic evaluation of content-based emotion recognition to identify a potential *glass ceiling* in the use of regression. 160 audio features were tested in four categories, timbral, loudness, harmonic, and rhythmic (with or without feature selection), as well as different regressors in three categories, Linear Regression, variants of regression trees and SVRs with Radial Basis Function (RBF) kernel (with or without parameter optimisation). Ground truth data were collected to indicate *induced* emotion, as in [69], by averaging arousal and valence scores from 50 subjects for 288 music pieces. Confirming earlier findings that arousal is easier to predict than valence, peak R^2 of 69.7% (arousal) and 25.8% (valence) were obtained using SVR-RBF. The authors concluded that small database size presents a major problem, while the wide distribution of individual responses to a song spreading in the AV space was seen as another limitation. In order to overcome the subjectivity and potential nonlinearity of AV coordinates collected from users, and to ease the cognitive load during data collection, Yang et al. proposed a method to automatically determine the AV coordinates of songs using pair-wise comparison of relative emotion differences between songs using a ranking algorithm [67]. They demonstrated that the increased reliability of ground truth pays off when different learning algorithms are compared. In [68], the authors modeled emotions as probability distributions in the AV space as opposed to discrete coordinates. They developed a method to predict these distributions using *regression fusion*, and reported a weighted R^2 score of 54.39%.

4.3 Methods for Music Emotion Variation Detection

It can easily be argued however that emotions are not necessarily constant during the course of a piece of music, especially in classical recordings. The problem of Music Emotion Variation Detection (MEVD) can be approached from two perspectives: the detection of time-varying emotion as a continuous trajectory in the AV space, or finding music segments that are correlated with well defined emotions. The task of dividing the music into several segments which contain homogeneous emotion expression was first proposed by Lu et al. [34]. In [70], the authors also proposed MEVD but by classifying features resulting from 10s segments with 33.3% overlap using a fuzzy approach, and then computing arousal and valence values from the fuzzy output vectors. Building on earlier studies, Schmidt et al. [50] demonstrated that emotion distributions may be modeled as 2D Gaussian distributions in the AV space, and then approached the problem of time-varying emotion tracking. In [50], they employed Kalman filtering in a linear dynamical system to capture the dynamics of emotions across time. While this method provided smoothed estimates over time, the authors concluded that the wide variance in emotion space dynamics could not be accommodated by the initial model, and subsequently moved on to use Conditional Random Fields (CRF), a probabilistic graphical model to approach the same problem [51]. In modeling complex emotion-space distributions as AV *heatmaps*, CRF outperformed the prediction of 2D Gaussians using MLR. However, the CRF model has higher computational cost.

4.4 Multi-Modal Approaches and Fusion Policies

The combination of multiple feature domains have become dominant in recent MER systems and a comprehensive overview of combining acoustic features with lyrics, social tags and images (e.g. album covers) is presented in [25]. In most works, the previously discussed machine learning techniques still prevail, however, different feature fusion policies may be applied ranging from concatenating normalised feature vectors (early fusion) to boosting, or ensemble methods combining the outputs of classifiers or regressors trained on different feature sets independently (late fusion). Late fusion is becoming dominant since it solves the issues related to tractability, and the curse of dimensionality affecting early fusion. Bischoff et al. [6] showed that classification performance can be improved by exploiting both audio features and collaborative user annotations. In this study, SVMs with RBF kernel outperformed logistic regression, random forest, GMM, K-NN, and decision trees in case of audio features, while the Naïve Bayes Multinomial classifier produced the best results in case of tag features. An experimentally-defined linear combination of the results then outperformed classifiers using individual feature domains. In a more recent study, Lin et al. [32] demonstrated that genre-based grouping complements the use of tags in a two-stage multi-label emotion classification system reporting an improvement of 55% when genre information was used. Finally, Schuller [55] et al. combined audio features with metadata and Web-mined lyrics. They used a stemmed bag of words approach to represent lyrics and editorial metadata, and also extracted mood concepts from lyrics using natural language processing. Ensembles of REPTrees (a variant of Decision Trees) are used in a set of regression experiments. When the domains were considered in isolation, the best performance was achieved using audio features (chords, rhythm, timbre), but taking into account all the modalities improved the results.

5 Discussion and Conclusions

The results from the audio mood classification (AMC) task ran at MIREX from 2007 to 2009, and that of studies published between 2009 and 2011 reviewed in this article, suggest the existence of a "glass ceiling" for MER at F-measure about 65%. In a recent study [45], high-level features (mode "majorness" and key "clarity") have shown to enhance emotion recognition in a more robust way than low-level features. In line with these results, we claim that in order to improve MER models, there is a need for new mid or high-level descriptors characterising musical clues, more adapted to *explain* our conditioning to musical emotions than low-level descriptors. Some of the findings in music perception and cognition [56], psycho-musicology [17] [18], and affective computing [39] have not yet been exploited or adapted to their full potential for music information retrieval. Most of the current approaches to emotion recognition articulate on black-box models which do not take into account the interpretability of the relationships between features and emotion components; this is a disadvantage when trying to understand the underlying mechanisms [64]. Other emotion representation models, the appraisal models [39], attempt to predict the association between

appraisal and emotion components making possible to interpret the relationships. Despite the promising applications of semantic web ontologies in the field of MIR, the ontology approach has only be scarcely used in MER. [62] proposed a music-mood specific ontology grounded in the Music Ontology [42], in order to develop a multi-modal MER model relying on audio content extraction and semantic association reasoning. Such approach is promising since the system from [62] achieved a performance increase of approximately 20% points (60.6%) in comparison with the system by Feng, Cheng and Yang (FCY1), proposed at MIREX 2009 [38]. Recent research focuses on the use of regression and attempt to estimate continuous-valued coordinates in emotion spaces, which may then be mapped to an emotion label or a broader category. The choice between regression and classification is however not straightforward, as both categorical and dimensional emotion models have strengths and weaknesses for specific applications. Retrieving labels or categories given the estimated coordinates is often necessary, which requires a mapping between the dimensional and categorical models. This may not be available for a given model, may not be valid from a psychological perspective, and may also be dependent on extra-musical circumstances. With regard to the use of multiple modalities, most studies to date confirm that the strongest factors enabling emotion recognition are indeed related to the audio content. However a glass ceiling seems to exist which may only be vanquished if both contextual features and features from different musical modalities are considered.

Acknowledgments. This work was partly funded by the TSB project 12033-76187 "Making Musical Mood Metadata" (TS/J002283/1).

References

1. Arnold, M.B.: Emotion and personality. Columbia University Press, New York (1960)
2. Asmus, E.P.: Nine affective dimensions (Test manual). Tech. rep., University of Miami (1986)
3. Banse, R., Scherer, K.R.: Acoustic profiles in vocal emotion expression. J. of Pers. and Social Psy. 70, 614–636 (1996)
4. Barrington, L., Turnbull, D., Yazdani, M., Lanckriet, G.: Combining audio content and social context for semantic music discovery. In: Proc. ACM SIGIR (2009)
5. Bischoff, K., Firan, C.S., Nejdl, W., Paiu, R.: Can all tags be used for search? In: Proc. ACM CIKM. pp. 193–202 (2008)
6. Bischoff, K., Firan, C.S., Paiu, R., Nejdl, W., Laurier, C., Sordo, M.: Music mood and theme classification - a hybrid approach. In: Proc. ISMIR. pp. 657–662 (2011)
7. Castellano, G., Caridakis, G., Camurri, A., Karpouzis, K., Volpe, G., Kollias, S.: Body gesture and facial expression analysis for automatic affect recognition, pp. 245–255. Oxford University Press, New York (2010)
8. Cowie, R., McKeown, G., Douglas-Cowie, E.: Tracing emotion: an overview. Int. J. of Synt. Emotions (2012)
9. Dang, T.T., Shirai, K.: Machine learning approaches for mood classification of songs toward music search engine. In: Proc. ICKSE (2009)
10. Davies, S., Allen, P., Mann, M., Cox, T.: Musical moods: a mass participation experiment for affective classification of music. In: Proc. ISMIR. pp. 741–746 (2011)
11. Eerola, T.: A comparison of the discrete and dimensional models of emotion in music. Psychol. of Mus. 39(1), 18–49 (2010)
12. Eerola, T., Lartillot, O., Toivianien, P.: Prediction of multidimensional emotional ratings in music from audio using multivariate regression models. In: Proc. ISMIR (2009)
13. Ekman, P., Friesen, W.V.: Facial Action Coding System. Consulting Psychologists Press, Palo Alto, CA (1978)
14. Essid, S., Richard, G., David, B.: Musical instrument recognition by pairwise classification strategies. IEEE Trans. on Audio, Speech, and Langu. Proc. 14(4), 1401–1412 (2006)

15. Feng, Y., Zhuang, Y., Pan, Y.: Popular music retrieval by detecting mood. Proc. ACM SIGIR pp. 375–376 (2003)
16. Fontaine, J.R., Scherer, K.R., Roesch, E.B., Ellsworth, P.: The world of emotions is not two-dimensional. Psychol. Sc. 18(2), 1050–1057 (2007)
17. Gabrielsson, A.: Emotional expression in synthesizer and sentograph performance. Psychomus. 14, 94–116 (1995)
18. Gabrielsson, A.: The influence of musical structure on emotional expression, pp. 223–248. Oxford University Press (2001)
19. Han, B.J., Dannenberg, R.B., Hwang, E.: SMERS: music emotion recognition using support vector regression. In: Proc. ISMIR. pp. 651–656 (2009)
20. Hevner, K.: Expression in music: a discussion of experimental studies and theories. Psychol. Rev. 42(2), 186–204 (1935)
21. Hevner, K.: Experimental studies of the elements of expression in music. Am. J. of Psychol. 48(2), 246–268 (1936)
22. Hu, X., Downie, J.S.: Exploring mood metadata: relationships with genre, artist and usage metadata. In: Proc. ISMIR (2007)
23. Huq, A., Bello, J.P., Rowe, R.: Automated music emotion recognition: A systematic evaluation. J. of New Mus. Res. 39(3), 227–244 (2010)
24. Kim, J.H., Lee, S., Kim, S.M., Yoo, W.Y.: Music mood classification model based on Arousal-Valence values. In: Proc. ICACT. pp. 292–295 (2011)
25. Kim, Y.E., Schmidt, E.M., Migneco, R., Morton, B.G.: Music emotion recognition: a state of the art review. Proc. ISMIR pp. 255–266 (2010)
26. Krumhansl, C.L.: An exploratory study of musical emotions and psychophysiology. Can. J. of Exp. Psychol. 51(4), 336–353 (1997)
27. Laukka, P., Elfenbein, H.A., Chui, W., Thingujam, N.S., Iraki, F.K., Rockstuhl, T., Althoff, J.: Presenting the VENEC corpus: Development of a cross-cultural corpus of vocal emotion expressions and a novel method of annotation emotion appraisals. In: Devillers, L., Schuller, B., Cowie, R., Douglas-Cowie, E., Batliner, A. (eds.) Proc. of LREC work. on Corp. for Res. on Emotion and Affect. pp. 53–57. European Language Resources Association, Paris (2010)
28. Lee, J.A., Downie, J.S.: Survey of music information needs, uses, and seeking behaviors: preliminary findings. In: Proc. ISMIR (2004)
29. Lee, S., Kim, J.H., Kim, S.M., Yoo, W.Y.: Smoodi: Mood-based music recommendation player. In: Proc. IEEE ICME. pp. 1–4 (2011)
30. Lesaffre, M., Leman, M., Martens, J.P.: A user oriented approach to music information retrieval. In: Proc. Content-Based Retriev. Conf. Daghstul Seminar Proceedings, Wadern Germany (2006)
31. Li, T., Ogihara, M.: Detecting emotion in music. Proc. ISMIR pp. 239–240 (2003)
32. Lin, Y.C., Yang, Y.H., Chen, H.H.: Exploiting online music tags for music emotion classification. ACM Trans. on Mult. Comp. Com. and App. 7S(1), 26:1–15 (2011)
33. Lin, Y.C., Yang, Y.H., Chen, H.H., Liao, I.B., Ho, Y.C.: Exploiting genre for music emotion classification. In: Proc. IEEE ICME. pp. 618–621 (2009)
34. Lu, L., Liu, D., Zhang, H.J.: Automatic mood detection and tracking of music audio signals. IEEE Trans. on Audio, Speech, and Langu. Proc. 14(1), 5–18 (2006)
35. Mann, M., Cox, T.J., Li, F.F.: Music mood classificationof television theme tunes. In: Proc. ISMIR. pp. 735–740 (2011)
36. McVicar, M., Freeman, T., De Bie, T.: Mining the correlation between lyrical and audio features and the emergence of mood. In: Proc. ISMIR. pp. 783–788 (2011)
37. Meyer, L.B.: Emotion and meaning in music. The University of Chicago press (1956)
38. MIREX: Audio mood classification (AMC) results. `http://www.music-ir.org/mirex/wiki/2009:Audio_Music_Mood_Classification_Results` (2009)
39. Mortillaro, M., Meuleman, B., Scherer, R.: Advocating a componential appraisal model to guide emotion recognition. Int. J. of Synt. Emotions (2012 (in press))
40. Myint, E.E.P., Pwint, M.: An approach for multi-label music mood classification. In: Proc. ICSPS. vol. VI, pp. 290–294 (2010)
41. Osgood, C.E., Suci, G.J., Tannenbaum, P.H.: The measurement of meaning. University of Illinois Press, Urbana (1957)
42. Raimond, Y., Abdallah, S., Sandler, M., Frederick, G.: The music ontology. In: Proc. ISMIR. Vienna, Austria (2007)
43. Roseman, I.J., Smith, C.A.: Appraisal theory: Overview, assumptions, varieties, controversies, pp. 3–19. Oxford University Press, New York (2001)
44. Russell, J.A.: A circumplex model of affect. J. of Pers. and Social Psy. 39(6), 1161–1178 (1980)
45. Saari, P., Eerola, T., Lartillot, O.: Generalizability and simplicity as criteria in feature selection: application to mood classification in music. IEEE Trans. on Audio, Speech, and Langu. Proc. 19(6), 1802–1812 (2011)
46. Sanden, C., Zhang, J.: An empirical study of multi-label classifiers for music tag annotation. Proc. ISMIR pp. 717–722 (2011)
47. Scherer, K.R., Brosch, T.: Culture-specific appraial biases contribute to emotion disposition. Europ. J. of Person. 288, 265–288 (2009)
48. Scherer, K.R., Schorr, A., Johnstone, T.: Appraisal processes in emotion: Theory, methods, research. Oxford University Press, New York (2001)

49. Schmidt, E.M., Kim, Y.E.: Prediction of time-varying musical mood distributions from audio. In: Proc. ISMIR. pp. 465–470 (2010)
50. Schmidt, E.M., Kim, Y.E.: Prediction of time-varying musical mood distributions using Kalman filtering. In: Proc. ICMLA. pp. 655–660 (2010)
51. Schmidt, E.M., Kim, Y.E.: Modeling musical emotion dynamics with conditional random fields. In: Proc. ISMIR. pp. 777–782 (2011)
52. Schmidt, E.M., Turnbull, D., Kim, Y.E.: Feature selection for content-based, time-varying musical emotion regression. In: Proc. ACM SIGMM MIR. pp. 267–273 (2010)
53. Schubert, E.: Measuring emotion continuously: Validity and reliability of the two-dimensional emotion-space. Austral. J. of Psychol. 51(3), 154–165 (1999)
54. Schubert, E.: Update of the Hevner adjective checklist. Percept. and Mot. Skil. pp. 117–1122 (2003)
55. Schuller, B., Weninger, F., Dorfner, J.: Multi-modal non-prototypical music mood analysis in continous space: reliability and performances. In: Proc. ISMIR. pp. 759–764 (2011)
56. Sloboda, J.A., Juslin, P.N.: Psychological perspectives on music and emotion, pp. 71–104. Series in Affective Science, Oxford University Press (2001)
57. Thayer, J.F.: Multiple indicators of affective responses to music. Dissert. Abst. Int. 47(12) (1986)
58. Tsunoo, E., Akase, T., Ono, N., Sagayama, S.: Music mood classification by rhythm and bass-line unit pattern analysis. In: Proc. ICASSP. pp. 265–268 (2010)
59. Vaizman, Y., Granot, R.Y., Lanckriet, G.: Modeling dynamic patterns for emotional content in music. In: Proc. ISMIR. pp. 747–752 (2011)
60. Vuoskoski, J.K.: Measuring music-induced emotion: A comparison of emotion models, personality biases, and intensity of experiences. Music. Sc. 15(2), 159–173 (2011)
61. Waletzky, J.: Bernard Hermann Music For the Movies. DVD Les Films d'Ici / Alternative Current (1992)
62. Wang, J., Anguerra, X., Chen, X., Yang, D.: Enriching music mood annotation by semantic association reasoning. In: Proc. Int. Conf. on Mult. (2010)
63. Wang, X., Chen, X., Yang, D., Wu, Y.: Music emotion classification of Chinese songs based on lyrics using TF*IDF and rhyme. In: Proc. ISMIR. pp. 765–770 (2011)
64. Wehrle, T., Scherer, K.R.: Toward computational modelling of appraisal theories, pp. 92–120. Oxford University Press, New York (2001)
65. Whissell, C.M.: The dictionary of affect in language, vol. 4, pp. 113–131. Academic Press, New York (1989)
66. Wieczorkowska, A., Synak, P., Ras, Z.W.: Multi-label classification of emotions in music. Proc. Intel. Info. Proc. and Web Min. pp. 307–315 (2006)
67. Yang, Y.H., Chen, H.H.: Ranking-based emotion recognition for music organisation and retrieval. IEEE Trans. on Audio, Speech, and Langu. Proc. 19(4), 762–774 (2010)
68. Yang, Y.H., Chen, H.H.: Prediction of the distribution of perceived music emotions using discrete samples. IEEE Trans. on Audio, Speech, and Langu. Proc. 19(7), 2184–2195 (2011)
69. Yang, Y.H., Lin, Y.C., Su, Y.F., Chen, H.H.: A regression approach to music emotion recognition. IEEE Trans. on Audio, Speech, and Langu. Proc. 16(2), 448–457 (2008)
70. Yang, Y.H., Liu, C.C., Chen, H.H.: Music emotion classification: A fuzzy approach. Proc. ACM Int. Conf. on Mult. pp. 81–84 (2006)
71. Zentner, M., Grandjean, D., Scherer, K.R.: Emotions evoked by the sound of music: Differentiation, classification, and measurement. Emotion 8(4), 494–521 (2008)
72. Zhao, Y., Yang, D., Chen, X.: Multi-modal music mood classification using co-training. In: Proc. Int. Conf. on Comp. Intel. and Soft. Eng. (CiSE). pp. 1–4 (2010)
73. Zhao, Z., Xie, L., Liu, J., Wu, W.: The analysis of mood taxonomy comparison between Chinese and Western music. In: Proc. ICSPS. vol. VI, pp. 606–610 (2010)

A Feature Survey for Emotion Classification of Western Popular Music

Scott Beveridge[1] and Don Knox[2]

[1] Fraunhofer Institute for Digital Media Technology, Ilmenau, Germany
bevest@idmt.fraunhofer.de
[2] Glasgow Caledonian University, Cowcaddens Road, Glasgow, Scotland
d.knox@gcu.ac.uk

Abstract. In this paper we propose a feature set for emotion classification of Western popular music. We show that by surveying a range of common feature extraction methods, a set of five features can model emotion with good accuracy. To evaluate the system we implement an independent feature evaluation paradigm aimed at testing the property of generalizability; the ability of a machine learning algorithm to maintain good performance over different data sets.

Keywords: Music emotion classification, popular music, support vector machine

1 Introduction

Developing computational models of musical emotions is a multidisciplinary task including the fields of music psychology, musicology and computer science. Early research in this area focussed primarily on the classical music repertoire (1; 2; 3). From a musicological perspective this bias is easy to understand. Classical content provides well structured and well defined emotional ideas by means of motif, movement and form. In comparison, popular music tends to be more sonically and emotionally homogeneous owing perhaps to the commercial nature of the genre. Nevertheless, it is important that in 'real world' applications of emotion classification this type of content is taken into account.

The aim of this paper is to identify a subset of musical features that characterizes emotion in Western popular music. It achieves this by examining six of the most commonly occurring feature extraction toolboxes in the Music Emotion Classification (MEC) literature. To test the robustness of this approach we adopt a number of feature selection and classification algorithms in the context of an independent feature evaluation paradigm. The objective is to examine model generalizability, the property of a machine learning model that ensures good performance over multiple unrelated data sets. Evidence of generalizability supports the universality of the selected features.

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

2 Feature Space

The feature spaces considered in this research are shown in Table 1. These algorithms extract low to mid-level acoustical and psychoacoustical features from the spectral representation of music clips. In all cases default parameters are used that include factors such as sample rate, bit depth, frame size and hop factor.

Toolbox	Number of features
MIRtoolbox	376
PsySound3	24
Marsyas 0.4	124
Marsyas 0.1	32
Sound Description Toolbox	187
Lu Implementation	71
All features	814

Table 1. Feature extraction toolboxes

Due to its demonstrated success in Music Emotion Recognition (MER) applications (4; 5) the MIRtoolbox for Matlab is used as an experimental baseline. The current version (1.3.4) provides a base set of 376 features derived from the statistics of frame-level features. PsySound3 creates features based on psychoacoustic models and is represented by 24 core features including those used in research by Yang *et al* (6). Marsyas, which was perhaps the first extraction framework to be developed for MIR, is implemented in two forms. The first version of the toolbox (0.1) contains a subset extractor which enjoyed success in early genre recognition tasks (7). The newest iteration of Marsyas (0.4) is also evaluated as it includes an extended feature extractor that is widely used in MIR. The Sound Description Toolbox has been included as it contains a number of MPEG-7 standard descriptors as well as perceptual and spectral features. Finally, the framework implemented in research by Lu *et al* (3) has been included. Although not publicly available this framework was recreated by the authors (8) due to its excellent performance in classical MER.

3 Emotion Space

Emotion concepts are defined on the basis of the circumplex model proposed by Russell (9). The circumplex model represents the emotion space with two orthogonal bipolar dimensions of *arousal* and *valence* (Figure 1). For the classification task the quadrants created by these dimensions are used as the target emotion concepts. These are anxious, exuberant, depressed and content.

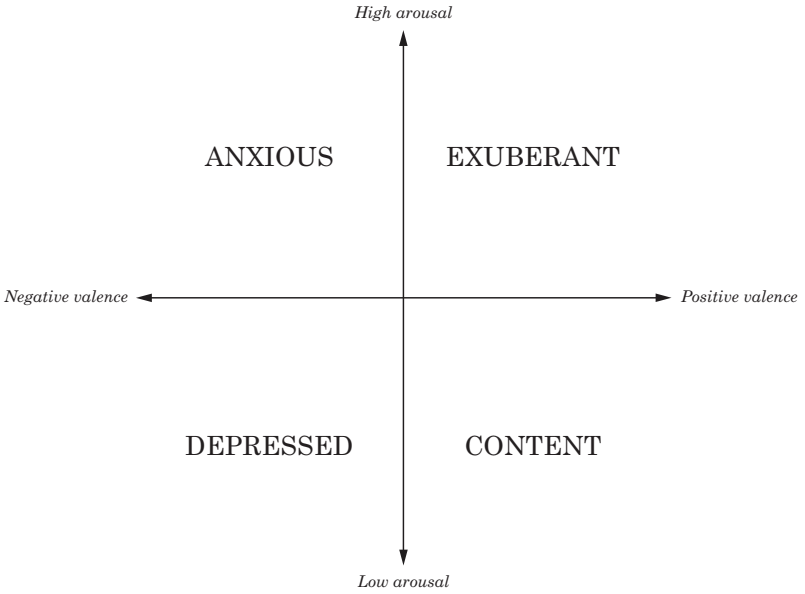


Fig. 1. The circumplex model

4 Musical Corpora

4.1 LastFM100

Two independent corpora are implemented in the classification framework. The first is derived from the LastFM database and consists of 25 tracks representing each quadrant of the circumplex model. These tracks were obtained by querying the LastFM database using the publicly available Application Programming Interface (API)[3]. This data acquisition technique has been used successfully in previous studies (10; 11) and is considered reliable due to its large number of active users.

4.2 Yang40

The second corpus was sourced from published research conduced by Yang *et al* (12). It contains 60 popular music tracks evaluated by 40 participants based on the dimensions arousal and valence. As this analysis requires discrete classes, each track was generalized into quadrants of the circumplex model determined by its arousal and valence values. Carrying out this process led to an uneven distribution across the emotion classes. To address this issue, random instances were removed from each class. This resulted in 10 instances or tracks per class.

[3] `www.last.fm/api`

5 Methodology

5.1 General Approach

A two-stage classification methodology was performed to determine the most representative feature set for Western popular music. First, initial models were constructed and evaluated in a traditional supervised train/test procedure. Using 10 x 10 fold stratified cross-validation these models were compared with respect to classification accuracy. In the second stage, the highest performing models were chosen for validation with the Yang40 corpus. As this data set is completely independent, this stage tests generalizability. High and consistent accuracy across data sets indicates high discriminative value of selected features. Each feature extraction toolbox was tested in isolation and then concatenated into a combined feature space named *Combined*.

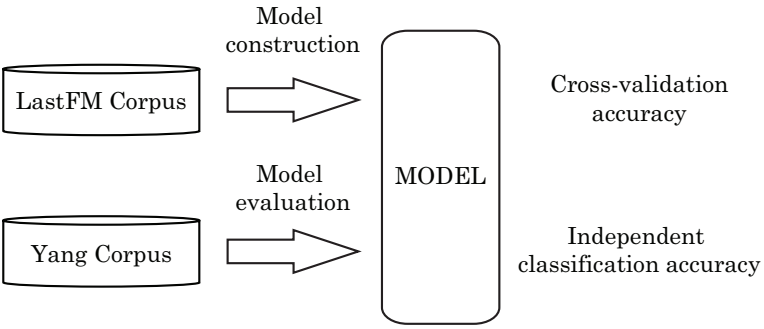


Fig. 2. Model training evaluation

5.2 Feature Selection and Classification

With such a high dimensional feature space it was necessary to apply feature reduction techniques. These included attribute subset and single attribute methods including InfoGainAttributeEval (InfoGain), CfsSubsetEval (Cfs), and ReliefFAttributeEval (ReliefF) (13). Based on the number of instances in the cross-validation data set and the need for parsimony, the 10 highest ranking features were chosen to represent the final feature space. Three classification models were also chosen for the analysis, K-Nearest Neighbours (K-NN), Naive Bayes and Support Vector Machines (SVM). A detailed explanation on the operation of these models is beyond the scope of this paper, however their inclusion was made based on good performance in previous emotion classification tasks (14; 15; 16; 17).

Performance of the feature selection/classification models are reported in terms of accuracy as defined by the Music Information Retrieval Evaluation eXchange (MIREX) in the 2007 Audio Music Mood Classification (AMC) task[4] (1).

$$\text{accuracy} = \frac{\text{number of correctly classified songs}}{\text{total number of songs}} \tag{1}$$

Modelling and feature selection were implemented in the WEKA environment, a freeware machine learning suite developed by the university of Waikato, New Zealand[5].

6 Results

6.1 Cross-validation

Table 2 shows classification accuracies for 10 runs of 10 fold cross-validation. For each feature extraction toolbox all feature selection/classification permutations are tested. The figures in bold show the highest performing models for each extraction toolbox including the concatenated *Combined* version.

These results show a clear boundary in performance between the Lu Implementation and *Combined* feature spaces and the rest of the feature toolboxes. The highest performing model for the *Combined* feature space has an accuracy of 0.64 with InfoGain feature selection and Naive Bayes classifier. The Lu Implementation is marginally higher with an accuracy of 0.65 using a combination of ReliefF feature selection and SVM classifier. The remaining toolboxes have accuracies ranging from 0.41 (Marsyas 0.1) to 0.48 (Sound Description Toolbox). This difference in performance is evident across all feature selection/classification approaches. As a result, the Lu Implementation modelled with SVM/ReliefF and *Combined* feature space modelled with Naive Bayes/InfoGain were chosen for independent validation with the Yang40 corpus.

6.2 Independent evaluation

The results of the independent evaluation step are shown in Table 3. With the 10 highest ranking features the Lu Implementation shows an accuracy of 0.65 and the *Combined* feature space 0.68. When expressed as percentages, this shows that 65 and 68% of the instances in the Yang40 corpus were correctly classified. As an additional measure the number of features used for classification were reduced to 5 and then 3. The aim was to determine how the steep drop in features might affect overall classification performance. Using only the top 5 ranked features, accuracies of 0.60 and 0.65 were achieved with the Lu Implementation and *Combined* feature spaces respectively. Using 3 features, classification accuracy dropped to 0.58 and 0.62. This small drop of between 6 and 7% shows the strong predictive power of these features.

[4] `http://www.music-ir.org/mirex/wiki/2007:Audio_Music_Mood_Classification`
[5] `http://www.cs.waikato.ac.nz/ml/weka/`

	InfoGain	ReliefF	Cfs
KNN			
MIR Toolbox	0.36	0.38	0.37
Marsyas 0.4	**0.45**	0.41	0.43
Marsyas 0.1	0.39	0.40	0.41
PsySound3	0.40	0.36	0.35
Sound Description Toolbox	0.39	0.41	0.44
Lu Implementation	0.60	0.62	0.63
Combined	0.63	0.64	0.57
Naive Bayes			
MIR Toolbox	0.41	**0.44**	0.40
Marsyas 0.4	0.40	0.40	0.43
Marsyas 0.1	**0.41**	0.41	0.40
PsySound3	0.41	0.42	0.40
Sound Description Toolbox	**0.48**	0.45	0.43
Lu Implementation	0.64	0.65	0.65
Combined	**0.64**	0.61	0.61
SVM			
MIR Toolbox	0.40	0.42	0.39
Marsyas 0.4	0.38	0.38	0.41
Marsyas 0.1	0.40	0.39	0.37
PsySound3	0.42	**0.43**	0.39
Sound Description Toolbox	0.44	0.45	0.43
Lu Implementation	0.62	**0.65**	0.62
Combined	0.62	0.59	0.58

Table 2. Model accuracies for 10 x 10 fold cross-validation

	Number of features		
Toolbox	10	5	3
Lu Implementation	0.65	0.60	0.58
Combined	0.68	0.65	0.62

Table 3. Classification Accuracy with 10, 5 and 3 features

7 Discussion

An important insight into the sonic properties of Western popular music is given in the reduced ranked feature set in Table 4. The two highest ranking features are statistics of frame-level values of spectral centroid. Indicating the 'centre of mass' of the spectrum, these features are a measure of high frequency content or brightness. Spectral flux, the third feature has been shown to be a useful perceptual indicator of music instrument timbre (18). The following two features relate to measures of intensity or loudness. These are Intensity ratio in sub band three as defined by Lu in (3), and sharpness, a perceptual measure of loudness relating to critical bandwidth. The sixth feature is a measure of tonal centre and is calculated as the mean of frame-wise centroid values of the chromagram. Spectral entropy is ranked seventh and gives an indication of the presence of predominant peaks in the signal. This is based on the Shannon entropy used in information theory (19). The next feature is Spectral Rolloff, an estimation of the amount of high frequency energy in a signal. The ninth ranked feature is Spectral Dissonance from the PsySound3 toolbox. Spectral Dissonance is a measure of the interference or *roughness* of spectral components. The final feature is the mean of the zerocross rate across frames. Zerocross is considered as a general measure of noisiness.

Overall, the ranking in Table 4 shows the importance of timbral characteristics in the recognition of emotion in popular music. The slight bias towards these features also suggests that modern production techniques, in particular over-compression, leads to homogeneity in terms of intensity or perceived loudness.

Rank	Feature name	Toolbox
1	Spectral Centroid Std	Lu Implementation
2	Spectral Centroid Variance	Lu Implementation
3	Spectral Flux Mean	Marsyas 0.1
4	Intensity Ratio Sub band 3 Mean	Lu Implementation
5	Sharpness Mean	PsySound3
6	Tonal Chromagram Centroid Mean	MIR Toolbox
7	Spectral Entropy Mean	MIR Toolbox
8	Spectral Rolloff Std	MIR Toolbox
9	Spectral Dissonance Std	PsySound3
10	Zerocross Mean	MIR Toolbox

Table 4. Ranked features from Combined feature space

8 Conclusions

By surveying six commonly used feature extraction toolboxes we present a compact feature subset for characterizing emotion in Western popular music. The efficacy of this feature space is tested using a combination of feature selection and classification algorithms in a unique feature evaluation paradigm. By examining model generalizability we have shown the potential universality of these features in an emotion classification task. The composition of the final feature set shows a bias towards spectrally derived acoustic features, reinforcing the idea that modern production techniques remove some important information carried by intensity or loudness features.

9 Acknowledgements

The research presented in this paper is part of the SyncGlobal project. SyncGlobal is a 2-year collaborative research project between Piranha Womex AG, Bach Technology GmbH, 4FreindsOnly AG and the Fraunhofer IDMT in Ilmenau, Germany. The project is co-financed by the Germany Ministry of Education and Research in the framework of the SME innovation program (FKZ 01/S11007).

Bibliography

[1] P. N. Juslin and J. A. Sloboda, *Music and Emotion: Theory and Research*. Oxford University Press, 2001.

[2] D. Liu, N. Zhang, and H. Zhu, "Form and mood recognition of johann strauss's waltz centos," *The Chinese Journal of Electronics*, vol. 12, no. 4, pp. 587–593, 2003.

[3] L. Lu, D. Liu, and H. J. Zhang, "Automatic mood detection and tracking of music audio signals," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 5–18, 2006.

[4] T. Eerola, O. Lartillot, and P. Toiviainen, "Prediction of multidimensional emotional ratings in music from audio using multivariate regression models," in *Proceedings of the 10th International Society for Music Information Retrieval (ISMIR) Conference*, pp. 621–626, 2009.

[5] J. C. Wang, H. Y. Lo, S. K. Jeng, and H. M. Wang, "MIREX 2010: Audio classification using semantic transformations and classifier ensemble," tech. rep., Institute of Information Science, Academia Sinica, Taipei, Taiwan, 2010. Available from URL `http://www.music-ir.org/mirex/abstracts/2010/WLJW2.pdf`, accessed January 2011.

[6] Y. H. Yang, Y. C. Lin, Y. F. Su, and H. H. Chen, "A regression approach to music emotion recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 448–457, 2008.

[7] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[8] D. Knox, S. Beveridge, L. Mitchell, and R. A. R. MacDonald, "Acoustic analysis and mood classification of pain-relieving music," *The Journal of the Acoustical Society of America*, vol. 130, no. 3, pp. 1673–1682, 2011.

[9] J. Russell, "A circumplex model of emotions," *The Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[10] X. Hu, M. Bay, and J. S. Downie, "Creating a simplified music mood classification ground-truth set," in *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR07)*, pp. 309–310, 2007.

[11] Y. C. Lin, Y. H. Yang, H. H. Chen, I. B. Liao, and Y. C. Ho, "Exploiting genre for music emotion classification," in *IEEE International Conference on Multimedia and Expo, (ICME 2009)*, pp. 618–621, 2009.

[12] Y. H. Yang, Y. F. Su, Y. C. Lin, and H. H. Chen, "Music emotion recognition: the role of individuality," in *Proceedings of the international workshop on Human-centered multimedia*, pp. 13–22, 2007.

[13] E. Frank and I. H. Witten, *Data Mining: Practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann, 2005.

[14] G. Tzanetakis and P. Cook, *Manipulation, analysis and retrieval systems for audio signals*. PhD thesis, 2002.

[15] D. Turnbull, L. Barrington, and G. Lanckriet, "Modelling music and words using a multi-class naive bayes approach," in *Proceedings of the 7th International Society for Music Information Retrieval (ISMIR) Conference*, 2006.

[16] K. Bischoff, C. Firan, R. Paiu, W. Nejdl, C. Laurier, and M. Sordo, "Music mood and theme classification-a hybrid approach," in *Proceedings of the International Society for Music Information Retrieval Conference, Kobe, Japan*, 2009.

[17] R. P. Panda, Renato; Paiva, "Using support vector machines for automatic mood tracking in audio music," in *Audio Engineering Society Convention 130*, 2011.

[18] S. Le Groux and P. Verschure, "Emotional responses to the perceptual dimensions of timbre: A pilot study using physically informed sound synthesis," in *Proceedings of the 7th International Symposium on Computer Music Modeling*, 2010.

[19] C. Shannon, "A mathematical theory of communication," *Bell Systems Technical Journal*, vol. 27, pp. 379–423, 1948.

Support Vector Machine Active Learning for Music Mood Tagging

Álvaro Sarasúa, Cyril Laurier and Perfecto Herrera

Music Technology Group, Universitat Pompeu Fabra
{alvarosarasua, cyril.laurier}@gmail.com, perfecto.herrera@upf.edu

Abstract. *Active learning* is a subfield of machine learning based on the idea that the accuracy of an algorithm can be improved with fewer training samples if it is allowed to choose the data from which it learns. We present the results for Support Vector Machine (SVM) active learning experiments for music mood tagging based on a multi-sample selection strategy that chooses samples according to their proximity to the boundary, their proximity to points in the training set and the density around them. The influence of those key active learning parameters is assessed by means of ANalysis Of Variance (ANOVA). Using these analyses we demonstrate the efficiency of active learning compared to typical full-dataset batch learning: our method allows to tag music by mood more efficiently than a regular approach, requiring fewer instances to obtain the same performance than using random sample selection methods.

Keywords: active learning, music mood detection, support vector machines

1 Introduction

Detection of moods and emotions in music is a topic of increasing interest in which many problems and issues are still to be explored. So far, most works have dealed with the so-called "basic emotions"such as happiness, sadness, anger, fear and disgust [1] [2]. This work tries to give one step towards detecting music emotions that are more specific and hard to articulate. One of the factors that make non-basic mood detection difficult is that they are usually perceived with less agreement among listeners than basic ones [1]. In such a context, user-tailored systems that learn from user perception become a must. Usually, such systems require getting a big amount of information from the user (e.g. using relevance feedback or other techniques) and his/her tastes in a process that can take too long.

In our case, we extend Laurier's method [3] for music mood classification. In such system, mood tags are assigned by means of a two-step process of feature extraction and statistical analysis. Those features are obtained for labeled songs and then the system is trained to learn which values of the extracted features define every group. A careful selection of the training examples is always required in order to maximize the generalization power of the final system.

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

Our work deals with the task of optimizing the training process by trying to speed user customization up, keeping the system as general as possible to different music genres for a specific user. We use **active learning** as it has shown good performance on many multimedia applications [4] [5] [6] but, surprisingly, it has been rarely used in Music Information Retrieval (MIR) even though its promising results [7] [8]. We perform a study on the application of active learning techniques to music mood classification extending Laurier's method with a multi-sample selection strategy based on Wang's [8]. In addition, we study the influence that all the parameters of this strategy have on the final results.

During this introduction, we review some concepts about active learning and its application to MIR. In Section 2 we explain our methodology. Results are shown and explained in Section 3. Finally, we discuss these results in Section 4.

1.1 Active Learning

Active Learning (AL) is aimed at maximizing the accuracy of a machine learning algorithm by means of allowing it to choose the data from which it learns (this usually leading to minimizing the size of the training set). In order to do so, the system may pose *queries* (unlabeled data instances) to be labeled by an *oracle*. AL is useful for problems where unlabeled data is easy to obtain, while labels are not [9]. We will deal with **uncertainty-based AL**, in which the learner queries instances for which it is least certain about how to label. The basic idea is that if, for example, the classification is binary, the instance that would be queried would be the one which probability of being positive is closest to 0.5 (total uncertainty). For more information about refinements of this technique we refer to [9].

Active Learning in MIR

Mandel et al. [7] demonstrated that AL techniques can be used for music retrieval with quite good results. Specifically, they classified songs according to their *style* and *mood*, being able to perform the same accuracy with half as many samples as without using AL to intelligently choose the training examples.

Wang *et al.* [8] propose a strategy for multi-samples selection for Support Vector Machine (SVM) AL. Their assumption is that, in music retrieval systems, it is necessary to present multiple samples (i.e. to make multiple queries) at each iteration. This is because the user could very likely lose patience after some time if just one sample is presented to him to label at each iteration. As we use the method they propose, we explain it in depth in Section 2.4.

2 Methodology

2.1 Mood tags and songs datasets

We used datasets from previous research ([3]) for *happy*, *sad*, *aggressive* and *relaxed* categories and we also created new datasets for *humorous*, *triumphant*,

mysterious and *sentimental* mood tags. These tags were chosen according to two main criteria: first, they improve the emotional representation given by the already existing ones (i.e. they do not have a close semantic relationship in the sense that they cover a broad and non-overlapping emotional landscape); second, they have a certain social relevance (judged by their presence as tags in lastfm[1]) that makes them interesting to study.

As we deal with binary classification, for each mood tag we actually need two collections of songs: one containing songs that *belong* to that category and another one with songs that *do not belong* to it. They contain 150 to 200 30-seconds MP3 files and we have tried to get a high coverage of genres and artists. They were created by collecting a high number of songs according to their mood annotation at lastfm and then validating them by 6 listeners, keeping just those songs which tag was agreed by at least 4 of them. For further details on this part of the work, please refer to [10].

2.2 Feature Extraction and Dataset Management

Descriptors of timbre [11], loudness, rhythm [12] and tonal characteristics [13] were computed and processed using MTG-internal libraries [14]. We compute statistics of these values (min, max, mean, variance). Then, we normalize descriptors and reduce dimensionality using Principal Component Analysis (PCA). The number of components that is kept is different depending on the size of the dataset ($\approx$ datasetsize/20). Finally, we compute the density around each point. This is one of the parameters that the AL strategy uses to measure the *informativeness* of each point [8]. This strategy is explained with more detail in Section 2.4.

2.3 Active Learning Experiments

For our experiments, we set a scenario in which we simulate the interaction with a user. We do so because our work tries to study a large set of combinations of parameters. Performing experiments with users would be time-intensive and attention-demanding, thus increasing the risk of getting wrong input. We follow (as [7]) these steps:

1. Randomly split the database into equal-sized training and test sets.
2. Select a random sample from the training set as the seed (the song for which the user is looking for songs with the same mood).
3. If we are in the first round, select $ITS - 1$ (Initial Train Size) samples plus the seed as the initial set for feedback. We choose them randomly. Otherwise, select EPI (Elements to add Per Iteration) samples according to the sample selection strategy (see details in 2.4).
4. According to the ground truth, automatically label the selected samples (simulating user relevance feedback). Add the labels to the labeled dataset and remove them from the training set.

[1] `http://www.last.fm`

5. Retrain the SVM model with the available dataset. Get precision and recall over the test dataset.
6. Repeat 1-6 100 times to avoid being biased by the selected seed and initial random-selected training set.

2.4 Active Learning Strategies under Study

Wang's Multi-Sample Selection Strategy

In this approach, introduced in [8], multiple samples are selected in a way that they are i) not just close to the boundary (most uncertain/informative samples), but also ii) representative of the underlying structure and iii) not redundant among them. To fulfill these three criteria, three values are calculated on every iteration for each point and different weights can be given to them: the **distance to the decision boundary**, the **distance diversity** and the **density** around the sample.

The first one is given by the own SVM classifier, which calculates the decision boundary ant tells the distance of each point to it. The diversity is calculated every time a new unlabeled sample x is selected as a candidate to be added to the current sample set S. It is defined as the minimum distance between samples in the current selected sample set S (the higher the diversity, the more scattered the set of samples are in space) and is calculated as

$$Diver(S+x) = \arg \min_{x_i, x_j \in \{S+x\}} D(x_i, x_j) \tag{1}$$

Where $D(x_i, x_j)$ is the distance between points x_i and x_j and can be calculated using the function $\Phi(x)$ that maps points into the transformed SVM space:

$$D(x_i, x_j) = \sqrt{(\Phi(x_i) - \Phi(x_j))^2} = \sqrt{\Phi(x_i)^2 + \Phi(x_j)^2 - 2 \cdot \Phi(x_i) * \Phi(x_j)} \tag{2}$$

Given that the kernel function $K(x, y)$ performs the dot product of two points in the transformed space, equation (2) can be rewritten as

$$D(x_i, x_j) = \sqrt{K(x_i, x_i) + K(x_j, x_j) - 2 \cdot K(x_i, x_j)} \tag{3}$$

The density around the sample, which is included to avoid choosing outliers as candidates, selects samples from the densest regions. An average distance $T(x)$ from a particular sample x to its 10 closest neighbors is computed off-line as

$$T(x) = \frac{D(x_{j1}, x) + \ldots D(x_{j10}, x)}{10}, x_{j1} \neq \ldots x_{j10} \tag{4}$$

Once these values are obtained, the selected point is the one that minimizes ($distance\_to\_boundary - diversity + density$) and the process is repeated as many times as samples are to be added at the current iteration.

Modified Wang's Multi-sample Selection Strategy

This strategy is proposed as an option to solve a drawback of uncertainty-based AL strategies which is even clearer when small training sets are used: reducing uncertainty may not always be the best choice. The idea is that the system should be quite sure about elements that fall far from the boundary, but if actually a newly retrieved tagged song of this kind is wrongly classified, the information it will bring to the system will be high as it will imply a big change on the decision boundary. Therefore, what we do is to perform exactly the same strategy just explained for half of the samples to be added, while the other half is actually selected from those furthest from the boundary.

3 Results

The experiment explained in Section 2.3 was performed for different sets of parameters for all the datasets. Although here we present some of the most relevant ones, please refer to [10] to find the results with all the possible considered combinations of *initial training size*, *elements per iteration*, combinations of weights for *distance to boundary*, *diversity* and *density* values and strategies (AL or random).

ANalysis Of VAriance (ANOVA) was used in order to test the influence of each of the parameters on one of the performance measures (F-measure). ANOVA looks for significant differences (i.e., unlikely to be found by chance) between means of different experimental conditions by comparing variances. The null hypothesis in the test is that the means of assumed normally distributed populations, all having the same standard deviation, are equal. In our specific case, each distribution corresponds to a certain configuration of values of one (several) parameter(s). If the null hypothesis is rejected, the value(s) of that (those) parameter(s) is considered to influence the results of F-measure.

ANOVA test were performed at different moments of the experiments for all the iterations. These tests showed that none of the parameters except the *Initial Training Size* had influence on the results for the first iteration, which is exactly what we could expect. Also, results show that the interactions between the method or combination of weights with *Initial Training Size* and *Elements Per Iteration* create significant differences in the F-measure mean. Another interesting observation is that using AL creates significant changes in the mean F-measure already by the second iteration.

Results in Table 1a confirm that there is an influence of METHOD (a variable coding random or each of the AL methods) on the results of the F-measure already in the second iteration, F (2, 115363 = 361.452), p = 000. As shown in Table 1b, the same applies for WEIGHTS (each value of WEIGHTS corresponds to a combination of the 3 weight values): it has a significant influence on the F-measure for the second iteration, F (3, 115354) = 195.085, p = 0.000. These results tell us that that changing the parameters and their combinations has an influence on the results that is not by chance.

Table 1a: Results of ANOVA test on the influence and interactions among *Initial Training Size(ITS)*, *Elements Per Iteration (EPI)* and *METHOD* on the F-measure in the first iteration.

Source	Degrees of freedom	F	p
METHOD	2	361.452	0.000
ITS	2	3959.617	0.000
EPI	2	993.995	0.000
METHOD*ITS	4	172.726	0.000
METHOD*EPI	4	60.673	0.000
ITS*EPI	4	185.998	0.000
METHOD*ITS*EPI	8	84.022	0.000
Error	115363		
Total	115390		

Table 1b: Results of ANOVA test on the influence and interactions among *Initial Training Size (ITS)*, *Elements Per Iteration (EPI)* and *WEIGHTS* on the F-measure in the second iteration. This table shows the influence of *WEIGHTS* on Wang's multi-sample selection strategy, therefore *METHOD* does not appear.

Source	Degrees of freedom	F	p
WEIGHTS	3	195.085	0.000
ITS	2	4498.762	0.000
EPI	2	1312.364	0.000
WEIGHTS*ITS	33.215	172.726	0.000
WEIGHTS*EPI	6	3.087	0.000
ITS*EPI	4	233.452	0.000
WEIGHTS*ITS*EPI	12	3.709	0.000
Error	115354		
Total	115390		

Figures 1a and 1b show average precision and recall values after 100 runs for both AL strategies and random sample selection for an initial training size of 5 samples, adding 8 elements per iteration and giving the same weight to the three parameters for the AL strategies. Precision and recall values are higher using Wang's multi-sample selection strategy (up to $\approx$ 4 percent point units higher precision than random sample selection at second iteration and $\approx$ 7 percent point units at third and fourth).

In Figures 1c and 1d each line corresponds to a different combination of weights for *distance to the boundary*, *diversity* and *density* using Wang's multi-sample selection strategy. The results show that the differences are not big, though the case in which *diversity* is given a higher weight is the one with the best performance (very close to the case in which all elements are given the same weight). The other two cases (higher weight for *distance to the boundary* or *density*) perform worse $\approx$ 5 percent point units lower precision in third iteration).

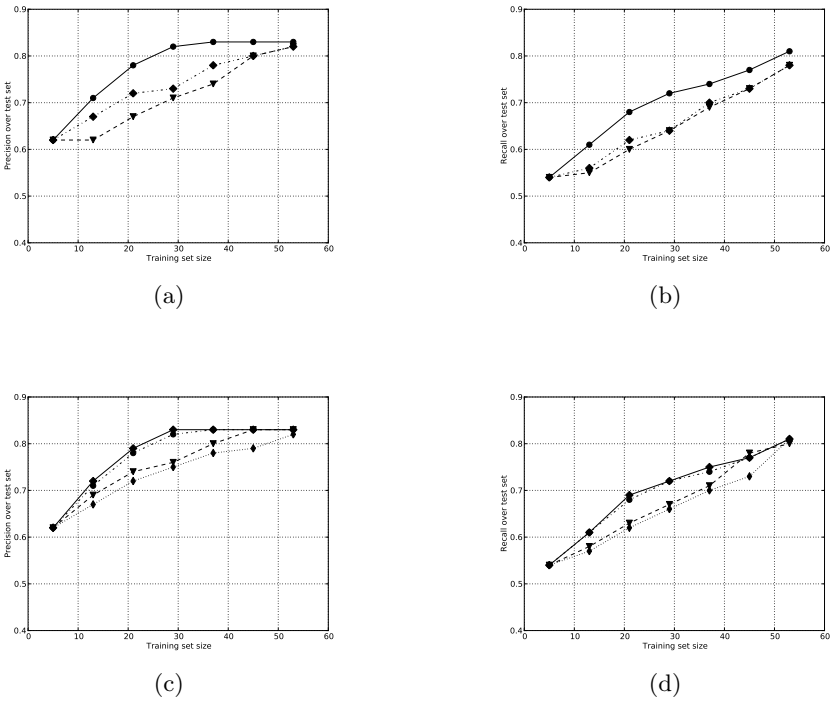


Fig. 1: (a) and (b) show precision and recall values during 7 rounds for different sample selection strategies.Wang's multi-sample selection strategy (•) converges faster faster to best performance than random sample selection (⋄) and modified Wang's strategy (▽). (c) and (d) show precision and recall values during 7 rounds for different weight combinations on Wang's multi-sample selection strategy. Best performance is achieved giving the same weight to the three parameters (•) or giving more weight to *diversity* (big ⋄). Results are worse for cases in which more weight is given to *distance to the boundary* (▽) or *density* (small ⋄).

4 Discussion and Future Work

Results of our experiments confirm those by Mandel [7] or Wang [8], in the sense that AL can help achieving a given performance on mood classification using less training instances than random sample selection. As shown in Fig. 1, the same precision can be achieved by means of AL with half instances than typical batch learning experiments. We also explored the influence of different parameters and determined that distance diversity plays a critical role for achieving good results. This is the parameter responsible for taking non-redundant samples, so one of our conclusions is that, in the presented scenario, it is more important to ensure learning about the whole distribution of points in the space rather than stressing other aspects. For example, by giving more weight to the distance to

the boundary, we may be querying for points that are too close to each other in every iteration, thus not learning about the whole dataset.

It may be interesting to explore different AL techniques. A comprehensive state-of-the-art review on AL can be found in [9], and we also present a brief review on these techniques in [10]. For example, Expected Error Reduction or Expected Variance Reduction AL techniques guarantee an improvement on the results, although they have a much higher computational cost.

Acknowledgments This work has been partially supported by the projects Classical Planet: TSI-070100- 2009-407 (MITYC) and DRIMS: TIN2009-14247-C02-01 (MICINN). We want to thank Nicolas Wack and Hendrik Purwins for their very valuable feedback.

References

1. Juslin, P.N., Sloboda, J.A.: Music and emotion: Theory and research. Oxford University Press Oxford, England (2001).
2. Ekman, P.: An argument for basic emotions. Cognition and Emotion. 6, 3, 169-200 (1992).
3. Laurier, C.: Automatic Classification of Musical Mood by Content Based Analysis. Universitat Pompeu Fabra, Barcelona (2011).
4. Settles, B., Craven, M.: An analysis of active learning strategies for sequence labeling tasks. Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP 08. 1070 (2008).
5. Tong, S., Koller, D.: Support Vector Machine Active Learning with Applications to Text Classification. Journal of Machine Learning Research. 45-66 (2001).
6. Chang, E.Y. et al.: Support Vector Machine Concept-Dependent Active Learning for Image Retrieval. IEEE Transactions on Multimedia. 1-35 (2005).
7. Mandel, M.I. et al.: Support vector machine active learning for music retrieval. Multimedia Systems. 12, 1, 3-13 (2006).
8. Wang, T.-J. et al.: Music retrieval based on a multi-samples selection strategy for support vector machine active learning. Proceedings of the 2009 ACM symposium on Applied Computing - SAC 09. 1750 (2009).
9. Settles, B.: Active learning literature survey. SciencesNew York. 15, 2, (2010).
10. Sarasúa, A.: Active Learning for User-Tailored Refined Music Mood Detection. Universitat Pompeu Fabra, Barcelona (2011).
11. Guaus, E.: Audio content processing for automatic music genre classification: descriptors, databases, and classifiers. Universitat Pompeu Fabra, Barcelona (2009).
12. Gouyon, F.: A Computation approach to rhythm description. Universitat Pompeu Fabra, Barcelona (2005).
13. Gómez, E.: Tonal description of music audio signals. Universitat Pompeu Fabra, Barcelona (2006).
14. Essentia & Gaia: audio analysis and music matching C++ libraries developed by the MTG (Resp.: Nicolas Wack), `http://mtg.upf.edu/technologies/essentia`

Modeling Expressed Emotions in Music using Pairwise Comparisons

Jens Madsen, Jens Brehm Nielsen, Bjørn Sand Jensen, and Jan Larsen *

Technical University of Denmark,
Department of Informatics and Mathematical Modeling,
Richard Petersens Plads B321, 2800 Lyngby, Denmark
{jenma; jenb; bjje; jl}@imm.dtu.dk

Abstract. We introduce a two-alternative forced-choice experimental paradigm to quantify expressed emotions in music using the two well-known arousal and valence (AV) dimensions. In order to produce AV scores from the pairwise comparisons and to visualize the locations of excerpts in the AV space, we introduce a flexible Gaussian process (GP) framework which learns from the pairwise comparisons directly. A novel dataset is used to evaluate the proposed framework and learning curves show that the proposed framework needs relative few comparisons in order to achieve satisfactory performance. This is further supported by visualizing the learned locations of excerpts in the AV space. Finally, by examining the predictive performance of the user-specific models we show the importance of modeling subjects individually due to significant subjective differences.

Keywords: expressed emotion, pairwise comparison, Gaussian process

1 Introduction

In recent years Music Emotion Recognition has gathered increasing attention within the Music Information Retrieval (MIR) community and is motivated by the possibility to recommend music that expresses a certain mood or emotion.

The design approach to automatically predict the expressed emotion in music has been to describe music by structural information such as audio features and/or lyrical features. Different models of emotion, e.g., categorical [1] or dimensional [2], have been chosen and depending on these, various approaches have been taken to gather emotional ground truth data [3]. When using dimensional models such as the well established *arousal* and *valence* (AV) model [2] the majority of approaches has been to use different variations of self-report direct scaling listening experiments [4].

* This work was supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886, and in part by the Danish Council for Strategic Research of the Danish Agency for Science Technology and Innovation under the CoSound project, case number 11-115328. This publication only reflects the authors' views.

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

Direct-scaling methods are fast ways of obtaining a large amount of data. However, the inherent subjective nature of both induced and expressed emotion, often makes anchors difficult to define and the use of them inappropriate due to risks of unexpected communication biases. These biases occur because users become uncertain about the meaning of scales, anchors or labels [5]. On the other hand, lack of anchors and reference points makes direct-scaling experiments susceptible to drift and inconsistent ratings. These effects are almost impossible to get rid of, but are rarely modeled directly. Instead, the issue is typically addressed through outlier removal or simply by averaging across users [6], thus neglecting individual user interpretation and user behavior in the assessment of expressed emotion in music.

Pairwise experiments eliminates the need for an absolute reference anchor, due to the embedded relative nature of pairwise comparisons which persists the relation to previous comparisons. However, pairwise experiments scale badly with the number of musical excerpts which they accommodate in [7] by a tournament based approach that limits the number of comparisons and transforms the pairwise judgments into possible rankings. Subsequently, they use the transformed rankings to model emotions.

In this paper, we present a novel dataset obtained by conducting a controlled pairwise experiment measuring expressed emotion in music on the dimensions of valence and arousal. In contrast to previous work, we learn from pairwise comparisons, directly, in a principled probabilistic manner using a flexible Gaussian process model which implies a latent but interpretable valence and arousal function. Using this latent function we visualize excerpts in a 2D valance and arousal space which is directly available from the principled modeling framework. Furthermore the framework accounts for inconsistent pairwise judgments by participants and their individual differences when quantifying the expressed emotion in music. We show that the framework needs relatively few comparisons in order to predict comparisons satisfactory, which is shown using computed learning curves. The learning curves show the misclassification error as a function of the number of (randomly chosen) pairwise comparisons.

2 Experiment

A listening experiment was conducted to obtain pairwise comparisons of expressed emotion in music using a two-alternative forced-choice paradigm. 20 different 15 second excerpts were chosen from the USPOP2002[1] dataset. The 20 excerpts were chosen such that a linear regression model developed in previous work [8] maps exactly 5 excerpts into each quadrant of the two dimensional AV space. A subjective evaluation was performed to verify that the emotional expression throughout each excerpt was considered constant.

A sound booth provided neutral surroundings for the experiment and the excerpts were played back using headphones to the 8 participants (2 female,

[1] `http://labrosa.ee.columbia.edu/projects/musicsim/uspop2002.html`

6 male). Written and verbal instructions were given prior to each session to ensure that subjects understood the purpose of the experiment and were familiar with the two emotional dimensions of valence and arousal. Each participant compared all 190 possible unique combinations. For the arousal dimension, participants were asked the question *Which sound clip was the most excited, active, awake?*. For the valence dimension the question was *Which sound clip was the most positive, glad, happy?*. The two dimensions were evaluated individually in random order. The details of the experiment are available in [9].

3 Pairwise-Observation based Regression

We aim to construct a model for the dataset given the audio excerpts in the set $\mathcal{X} = \{\mathbf{x}_i | i = 1, ..., n\}$ with $n = 20$ distinct excerpts, each described by an input vector $\mathbf{x}_i$ of audio features extracted from the excerpt. For each test subject the dataset comprises of all $m = 190$ combinations of pairwise comparisons between any two distinct excerpts, u and v, where $\mathbf{x}_u \in \mathcal{X}$ and $\mathbf{x}_v \in \mathcal{X}$. Formally, we denote the output set (for each subject) as $\mathcal{Y} = \{(d_k; u_k, v_k) | k = 1, ..., m\}$, where $d_k \in \{-1, 1\}$ indicates which of the two excerpts that had the highest valence or arousal. $d_k = -1$ means that the u_k'th excerpt is picked over the v_k'th and visa versa when $d_k = 1$.

We model the pairwise choice, d_k, between two distinct excerpts, u and v, as a function of the difference between two functional values, $f(\mathbf{x}_u)$ and $f(\mathbf{x}_v)$. The function $f : \mathcal{X} \to \mathbb{R}$ thereby defines an internal, but latent absolute reference of either valence or arousal as a function of the excerpt represented by the audio features.

Given a function, $f(\cdot)$, we can define the likelihood of observing the choice d_k directly as the conditional distribution.

$$p\left(d_k | \mathbf{f}_k\right) = \Phi\left(d_k \frac{f\left(\mathbf{x}_{v_k}\right) - f\left(\mathbf{x}_{u_k}\right)}{\sqrt{2}}\right), \tag{1}$$

where $\Phi(x)$ is the cumulative Gaussian (with zero mean and unity variance) and $\mathbf{f}_k = \left[f\left(\mathbf{x}_{u_k}\right), f\left(\mathbf{x}_{v_k}\right)\right]^\top$. This classical choice model can be dated back to Thurstone and his fundamental definition of *The Law of Comparative Judgment* [10].

We consider the likelihood in a Bayesian setting such that $p\left(\mathbf{f}|\mathcal{Y}, \mathcal{X}\right) = p\left(\mathcal{Y}|\mathbf{f}\right) p(\mathbf{f}|\mathcal{X}) / p\left(\mathcal{Y}|\mathcal{X}\right)$ where we assume that the likelihood factorizes, i.e., $p\left(\mathcal{Y}|\mathbf{f}\right) = \prod_{k=1}^{m} p\left(d_k|\mathbf{f}_k\right)$.

In this work we consider a specific prior, namely a Gaussian Process (GP), first considered with the pairwise likelihood in [11]. A GP is typically defined as "a collection of random variables, any finite number of which have a joint Gaussian distribution" [12]. By $f\left(\mathbf{x}\right) \sim \mathcal{GP}\left(\mathbf{0}, k(\mathbf{x}, \mathbf{x}')\right)$ we denote that the function $f(\mathbf{x})$ is modeled by a zero-mean GP with covariance function $k(\mathbf{x}, \mathbf{x}')$. The fundamental consequence of this formulation is that the GP can be considered a distribution over functions, defined as $p\left(\mathbf{f}|\mathcal{X}\right) = \mathcal{N}\left(\mathbf{0}, \mathbf{K}\right)$ for any finite set of of function values $\mathbf{f} = [f(\mathbf{x}_1), ..., f(\mathbf{x}_n)]^\top$, where $[\mathbf{K}]_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$.

Bayes relation leads directly to the posterior distribution over $\mathbf{f}$, which is not analytical tractable. Instead, we use the *Laplace Approximation* to approximate the posterior with a multivariate Gaussian distribution[1].

To predict the pairwise choice d_t on an unseen comparison between excerpts r and s, where $\mathbf{x}_r, \mathbf{x}_s \in \mathcal{X}$, we first consider the predictive distribution of $f(\mathbf{x}_r)$ and $f(\mathbf{x}_s)$. Given the GP, we can write the joint distribution between $\mathbf{f} \sim p(\mathbf{f}|\mathcal{Y}, \mathcal{X})$ and the test variables $\mathbf{f}_t = [f(\mathbf{x}_r), f(\mathbf{x}_s)]^T$ as

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_t \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{K} & \mathbf{k}_t \\ \mathbf{k}_t^T & \mathbf{K}_t \end{bmatrix} \right), \tag{2}$$

where $\mathbf{k}_t$ is a matrix with elements $[\mathbf{k}_t]_{i,2} = k(\mathbf{x}_i, \mathbf{x}_s)$ and $[\mathbf{k}_t]_{i,1} = k(\mathbf{x}_i, \mathbf{x}_r)$ with $\mathbf{x}_i$ being a training input.

The conditional $p(\mathbf{f}_t|\mathbf{f})$ is directly available from Eq. (2) as a Gaussian too. The predictive distribution is given as $p(\mathbf{f}_t|\mathcal{Y}, \mathcal{X}) = \int p(\mathbf{f}_t|\mathbf{f})\, p(\mathbf{f}|\mathcal{Y}, \mathcal{X})\, d\mathbf{f}$, and with the posterior approximated with the Gaussian from the Laplace approximation then $p(\mathbf{f}_t|\mathcal{Y}, \mathcal{X})$ will also be Gaussian given by $\mathcal{N}(\mathbf{f}_t|\boldsymbol{\mu}^*, \mathbf{K}^*)$ with $\boldsymbol{\mu}^* = \mathbf{k}_t^T \mathbf{K}^{-1} \hat{\mathbf{f}}$ and $\mathbf{K}^* = \mathbf{K}_t - \mathbf{k}_t^T (\mathbf{I} + \mathbf{W}\mathbf{K})\, \mathbf{k}_t$, where $\hat{\mathbf{f}}$ and $\mathbf{W}$ are obtained from the Laplace approximation (see [13]). In this paper we are only interested in the binary choice d_t, which is determined by which of $f(\mathbf{x}_r)$ or $f(\mathbf{x}_s)$ that dominates[2].

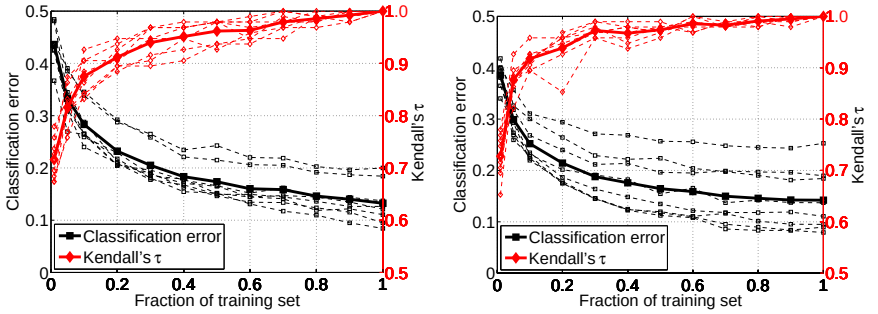
The zero-mean GP is fully defined by the covariance function, $k(\mathbf{x}, \mathbf{x}')$. In the emotion dataset each input instance is an excerpt described by the vector $\mathbf{x}$ containing the audio features for each time frame which is naturally modeled with a probability density, $p(\mathbf{x})$. We apply the probability product (PP) kernel [14] in order to support these types of distributional inputs. The PP kernel is defined directly as an inner product as $k(\mathbf{x}, \mathbf{x}') = \int [p(\mathbf{x})\, p(\mathbf{x}')]^q d\mathbf{x}$. We fix $q = 1/2$, leading to the Hellinger divergence [14]. In order to model the audio feature distribution for each excerpt, we resort to a (finite) Gaussian Mixture Model (GMM). Hence, $p(\mathbf{x})$ is given by $p(\mathbf{x}) = \sum_{z=1}^{N_z} p(z)\, p(\mathbf{x}|z)$, where $p(\mathbf{x}|z) = \mathcal{N}(\mathbf{x}|\mu_z, \sigma_z)$ is a standard Gaussian distribution. The kernel is expressed in closed form [14] as $k(p(\mathbf{x}), p(\mathbf{x}')) = \sum_z \sum_{z'} (p(z)\, p(z'))^q \tilde{k}(p(\mathbf{x}|\theta_z), p(\mathbf{x}'|\theta_{z'}))$ where $\tilde{k}(p(\mathbf{x}|\theta_z), p(\mathbf{x}'|\theta_{z'}))$ is the probability product kernel between two single components - also available in closed form [14].

4 Modeling Expressed Emotion

In this section we evaluate the ability of the proposed framework to capture the underlying structure of expressed emotions based on pairwise comparisons, directly. We apply the GP model using the probability product (PP) kernel described in Section 3 with the inputs based on a set of audio features extracted

[1] More details can be found in e.g. [13].

[2] With the pairwise GP model the predictive distribution of d_t can also be computed analytically (see [13]) and used to express the uncertainty in the prediction relevant for e.g. sequential designs, reject regions etc.



(a) Valence (b) Arousal

Fig. 1. Classification error learning curves and Kendall's τ for 10-fold CV on comparisons. Bold lines are mean curves across subjects and dash lines are curves for individual subjects. Notice, that for the classification error learning curves, the baseline performance corresponds to an error of 0.5, obtained by simply randomly guessing the pairwise outcome.

from the 20 excerpts. By investigating various combinations of features we obtained the best performance using two sets of commonly used audio features. The first set is the Mel-frequency cepstral coefficients (MFCC), which describe the short-term power spectrum of the signal. Secondly, we included spectral contrast features and features describing the spectrum of the Hanning windowed audio. Based on an initial evaluation, we fix the number of components in the GMM used in the PP Kernel to $N_z = 3$ components and train the individual GMMs by a standard EM algorithm with K-means initialization. Alternatively, measures such as the Bayesian Information Criterion (BIC) could be used to objectively set the model complexity for each excerpt.

4.1 Results: Learning Curves

Learning curves for the individual subjects are computed using 10-fold cross validation (CV) in which a fraction (90%) of the total number of pairwise comparisons constitutes the complete training set. Each point on the learning curve is an average over 10 randomly chosen and equally-sized subsets from the complete training set. The Kendall's τ rank correlation coefficient is computed in order to relate our results to that of e.g. [7] and other typical ranking based applications. The Kendall's τ is a measure of correlation between rankings and is defined as $\tau = (N_s - N_d)/N_t$ where N_s is the number of correctly ranked pairs, N_d is the number of incorrectly ranked pairs and N_t is the total number of pairs. The reported Kendall's τ is in all cases calculated with respect to the predicted ranks using all the excerpts.

Figure 1 displays the computed learning curves. With the entire training set included the mean classification errors across subjects for valence and arousal are 0.13 and 0.14, respectively. On average this corresponds to a misclassified comparison in every 7.5 and 7'th comparison for valence and arousal, respectively.

For valence, the mean classification error across users is below 0.2 with 40% of the training data included, whereas only 30% of the training data is needed to obtain similar performance for arousal. This indicates that the model for arousal can be learned slightly faster than valence. Using 30% of the training data the Kendall's τ is 0.94 and 0.97, respectively, indicating a good ranking performance using only a fraction of the training data.

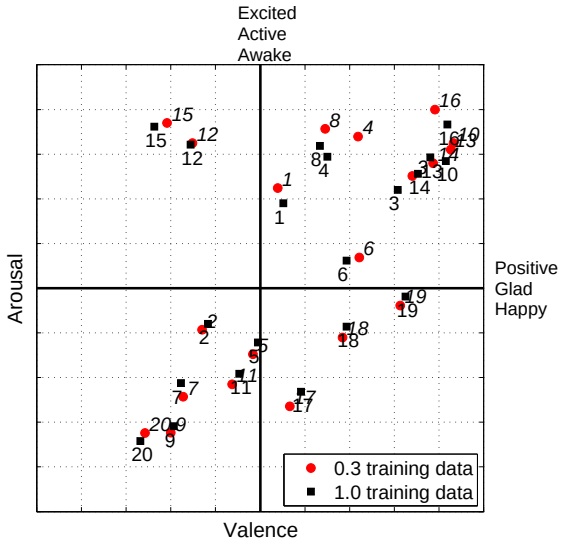
When considering the learning curves for individual users we notice significant individual differences between users—especially for arousal. Using the entire training set in the arousal experiment, the user for which the model performs best results in an error of 0.08 whereas the worst results in an error of 0.25. In the valence experiment the best and worst performances result in classification errors of 0.08 and 0.2, respectively.

4.2 Results: AV space

The learning curves show the pure predictive power of the model on unseen comparisons, but may be difficult to interpret in terms of the typical AV space. To address this we show that the latent regression function $f(\cdot)$ provides an internal but unit free representation of the AV scores. The only step required is a normalization which ensures that the latent values are comparable across folds and subjects. In Figure 2 the predicted AV scores are shown when the entire training set is included and when only 30% is included. The latter corresponds to 51 comparisons in total or an average of 2.5 comparisons per excerpt. The results are summarized by averaging across the predicted values for each user. 15 of the 20 excerpts are positioned in the typical high-valence high-arousal and low-valence low-arousal quadrants, 2 excerpts are clearly in the low-valance high-arousal quadrant and 3 excerpts are in the high-valance low-arousal quadrant of the AV space. The minor difference in predictive performance between 30% and the entire training dataset does not lead to any significant change in AV scores, which is in line with the reported Kendall's τ measure.

4.3 Discussion

The results clearly indicate that it is possible to model expressed emotions in music by directly modeling pairwise comparisons in the proposed Gaussian process framework using subject specific models. An interesting point is the large difference in predictive performance between subjects given the specific models. These differences can be attributed to the specific model choice (including kernel) or simply to subject inconsistency in the pairwise decisions. The less impressive predictive performance for certain subjects is presumably a combination of the two effects, although given the very flexible nature of the Gaussian process model, we mainly attribute the effect to subjects being inconsistent due to for example mental drift. Hence, individual user behavior, consistency and discriminative ability are important aspects of modeling expressed emotion in music and other cognitive experiments, and thus also a critical part when aggregating subjects in large datasets.



No.	Song name
1	311 - T and p combo
2	A-Ha - Living a boys adventure
3	Abba - Thats me
4	Acdc - What do you do for money honey
5	Aaliyah - The one i gave my heart to
6	Aerosmith - Mother popcorn
7	Alanis Morissette - These r the thoughts
8	Alice Cooper - Im your gun
9	Alice in Chains - Killer is me
10	Aretha Franklin - A change
11	Moby - Everloving
12	Rammstein - Feuer frei
13	Santana - Maria caracoles
14	Stevie Wonder - Another star
15	Tool - Hooker with a pen..
16	Toto - We made it
17	Tricky - Your name
18	U2 - Babyface
19	Ub40 - Version girl
20	Zz top - Hot blue and righteous

Fig. 2. AV values computed by averaging the latent function across folds and repetitions and normalizing for each individual model for each participant. Red circles: 30% of training set is used. Black squares: entire training set is used.

The flexibility and interpolation abilities of Gaussian Processes allow the number of comparisons to be significantly lower than the otherwise quadratic scaling of unique comparisons. This aspect and the overall performance should of course be examined further by considering a large scale dataset and the use of several model variations. In addition, the learning rates can be improved by combining the pairwise approach with active learning or sequential design methods, which in turn select only pairwise comparisons that maximize some information criterion.

We plan to investigate how to apply multi-task (MT) or transfer learning to the special case of pairwise comparisons, such that we learn one unifying model taking subjects differences into account instead of multiple independent subject-specific models. A very appealing method is to include MT learning in the kernel of the GP [15], but this might not be directly applicable in the pairwise case.

5 Conclusion

We introduced a two-alternative forced-choice experimental paradigm for quantifying expressed emotions in music in the typical arousal and valance (AV) dimensions. We proposed a flexible probabilistic Gaussian process framework to model the latent AV scales directly from the pairwise comparisons. The framework was evaluated on a novel dataset and resulted in promising error rates for both arousal and valence using as little as 30% of the training set corresponding to 2.5 comparisons per excerpt. We visualized AV scores in the well-known two dimensional AV space by exploiting the latent function in the Gaussian process

model, showing the application of the model in a standard scenario. Finally we especially draw attention to the importance of maintaining individual models for subjects due to the apparent inconsistency of certain subjects and general subject differences.

References

1. K. Hevner, "Experimental studies of the elements of expression in music," *American journal of Psychology*, vol. 48, no. 2, pp. 246–268, 1936.
2. J.A. Russell, "A circumplex model of affect," *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161, 1980.
3. Y.E. Kim, E.M. Schmidt, Raymond Migneco, B.G. Morton, Patrick Richardson, Jeffrey Scott, J.A. Speck, and Douglas Turnbull, "Music emotion recognition: A state of the art review," in *Proc. of the 11th Intl. Society for Music Information Retrieval (ISMIR) Conf*, 2010, pp. 255–266.
4. E. Schubert, *Measurement and time series analysis of emotion in music*, Ph.D. thesis, University of New South Wales, 1999.
5. M. Zentner and T. Eerola, *Handbook of Music and Emotion - Theory, Research, Application*, chapter 8 - Self-report measures and models, Oxford University Press, 2010.
6. A. Huq, J. P. Bello, and R. Rowe, "Automated Music Emotion Recognition: A Systematic Evaluation," *Journal of New Music Research*, vol. 39, no. 3, pp. 227–244, Sept. 2010.
7. Y.-H. Yang and H.H. Chen, "Ranking-Based Emotion Recognition for Music Organization and Retrieval," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 762–774, May 2011.
8. J. Madsen, *Modeling of Emotions expressed in Music using Audio features*, DTU Informatics, Master Thesis, http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6036, 2011.
9. J. Madsen, *Experimental Protocol for Modelling Expressed Emotion in Music*, DTU Informatics, http://www2.imm.dtu.dk/pubdb/p.php?6246, 2012.
10. L. L. Thurstone, "A law of comparative judgement.," *Psychological Review*, vol. 34, 1927.
11. W. Chu and Z. Ghahramani, "Preference learning with Gaussian Processes," *ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning*, pp. 137–144, 2005.
12. C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.
13. B.S. Jensen and J.B. Nielsen, *Pairwise Judgements and Absolute Ratings with Gaussian Process Priors*, Technical Report, DTU Informatics, http://www2.imm.dtu.dk/pubdb/p.php?6151, September 2011.
14. T. Jebara and A. Howard, "Probability Product Kernels," *Journal of Machine Learning Research*, vol. 5, pp. 819–844, 2004.
15. E.V. Bonilla, F.V. Agakov, and C.K.I. Williams, "Kernel multi-task learning using task-specific features," *Proceedings of the 11th AISTATS*, 2007.

Relating Perceptual and Feature Space Invariances in Music Emotion Recognition

Erik M. Schmidt, Matthew Prockup, Jeffery Scott, Brian Dolhansky, Brandon G. Morton, and Youngmoo E. Kim

Music and Entertainment Technology Laboratory (MET-lab)
Electrical and Computer Engineering, Drexel University
{eschmidt,mprockup,jjscott,bdol,bmorton,ykim}@drexel.edu

Abstract. It is natural for people to organize music in terms of its emotional associations, but while this task is a natural process for humans, quantifying it empirically proves to be a very difficult task. Consequently, no particular acoustic feature has emerged as the optimal representation for musical emotion recognition. Due to the subjective nature of emotion, determining how informative an acoustic feature domain is requires evaluation by human subjects. In this work, we seek to perceptually evaluate two of the most commonly used features in music information retrieval: mel-frequency cepstral coefficients and the chromagram. Furthermore, to identify emotion-informative feature domains, we seek to identify what musical features are most variant or invariant to changes in musical qualities. This information could also potentially be used to inform methods that seek to learn acoustic representations that are specifically optimized for prediction of emotion.

Keywords: emotion, music emotion recognition, features, acoustic features, machine learning, invariance

1 Introduction

The problem of automated recognition of emotional (or mood) content within music has been the subject of increasing attention among the music information retrieval (Music-IR) research community [1]. While there has been much progress in machine learning systems for estimating human emotional response to music, very little progress has been made in terms of compact or intuitive feature representations. Current methods generally focus on combining several feature domains (e.g. loudness, timbre, harmony, rhythm), in some cases as many as possible, and performing dimensionality reduction techniques such as principal component analysis (PCA). Overall, these methods have not sufficiently improved performance, and have done little to advance the field.

In this work, we begin by perceptually evaluating two of the most commonly used features in Music-IR: mel-frequency cepstral coefficients (MFCCs) and the chromagram. MFCCs have been shown in previous work to be one of the most informative feature domains for music emotion recognition [2–5], but as MFCCs

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

were originally designed for speech recognition, it is unclear why they perform so well or how much information about emotion they actually contain. Conversely, the chromagram appears to be one of the most intuitive representations, as it provides information about the notes contained in the piece, which could potentially provide information about the key and mode. Thus far, chroma has shown little promise in informing this problem. In order to properly assess these features, we construct a perceptual study using Amazon's Mechanical Turk[1] (MTurk) to analyze the relative emotion of two song clips, comparing human ratings of both the original audio and audio reconstructions from these features. By analyzing these reconstructions, we seek to directly assess how much information about musical emotion is retained in these features.

Given our collected data, we also wish to identify patterns in relationships between musical parameters (e.g. key, mode, tempo) and perceived emotion. By identifying variability in emotion related to these parameters, we identify existing features that respond with the highest variance to those that inform emotion, and the least variance in those that do not. In order to properly assess a large variety of features, we investigate the features used in our perceptual study reconstructions, features used in our prior work [2–5], and 14 additional features from the MIR-toolbox[2].

In investigating these invariances, we explore approaches that attempt to develop feature representations which are specifically optimized for the prediction of emotion. In forming such representations, we are presented with a very challenging problem as music theory offers an insufficient foundation for constructing features using a bottom-up approach. As a result, in previous work we have instead taken a top-down approach, attempting to learn representations directly from magnitude spectra [5]. These approaches show much promise but are highly underconstrained as we have little idea of what our features should be invariant to. In this paper, we seek to provide some initial insight into how these problems could be better constrained.

2 Background

A musical piece is made up of a combination of different attributes such as key, mode, tempo, instrumentation, etc. While not one of these attributes fully describes a piece of music, each one contributes to the listener's perception of the piece. We hope to establish which compositional attributes significantly determine emotion and which parameters are less relevant. These parameters are not the sole contributors to the emotion of the music, but are within our ability to measure from the symbolic dataset we use in our experiments, and therefore are the focus of this study [6]. Specifically, we want to determine whether these compositional building blocks induce changes in the acoustic feature domain.

[1] http://mturk.com

[2] http://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox

We motivate our experiments from findings that have been verified by several independent experiments in psychology [7–9]. When discussing emotion, we refer to happy versus sad temperament as valence and higher and lower intensity of that temperament as arousal [10]. Mode and tempo have been shown to consistently elicit a change in perceived emotion in user studies. Mode is the selection of notes (scale) that form the basic tonal substance of a composition and tempo is the speed of a composition [11]. Research shows that major modes tend to elicit happier emotional responses, while the inverse is true for minor modes [9, 12–14]. Tempo also determines a user's perception of music, with higher tempi generally inducing stronger positive valence and arousal responses [8, 9, 12, 13, 15].

3 Data Collection

In previous studies (such as [9]), several controlled variations of musical phrases are provided to each participant. Since we are studying the changes in the acoustic feature domain, we require samples that we can easily manipulate in terms of mode and tempo and that provide a wide enough range to ensure we are accurately representing all possible variations in the feature space. To this end, we put together a dataset of 50 Beatles MIDI files, attained online[3], spanning 5 albums (Sgt. Peppers, Revolver, Let It Be, Rubber Soul, Magical Mystery Tour). In order to remove the effect of instrumentation, each song was synthesized as a piano reduction and a random twenty second clip of each song was used for our labeling task.

3.1 Song Clip Pair Selection

Labeling the entire 1225 possible pairs from the 50 songs would be prohibitive so we choose to generate a subset of 160 pairs. Since the Beatles dataset we use contains 35 songs in the major (Ionian) mode and only 9 in the minor (Aolean) mode (with 6 additional pieces in alternate modes), we want to ensure that major-major pairings do not completely dominate our task. Some songs are represented one extra time in order to generate 160 pairs but no song is repeated more than once. Out of these 160 pairs, there are 81 major-major pairings, 33 major-minor pairings, and 7 minor-minor pairings.

For each song, we render the piano reduction of the MIDI file for the 20 second clip, and then compute MFCC and chroma features on the audio. After computing the features, we then synthesize audio from the features. Chromagram features are extracted and reconstructed using Dan Ellis' chroma features analysis and synthesis code[4] and MFCCs using his rastamat[5] library. The MFCC

[3] http://earlybeatles.com/

[4] http://www.ee.columbia.edu/~dpwe/resources/matlab/chroma-ansyn/

[5] http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/

reconstructions sound like a pitched noise source, and the chroma reconstructions have an ethereal 'warbly' quality to them but sound more like the original audio than the MFCC reconstruction (examples are available online [6]).

3.2 Mechanical Turk Annotation Task

In order to annotate our clip pairs, we use the Mechanical Turk online crowdsourcing engine to gain input from a wide variety of subjects [16]. In our Human Intelligence Task (HIT), we ask participants to label four uniformly selected song pairs from each of the three categories: original MIDI rendering, MFCC reconstructions, and chromagram reconstructions. For each pair of clips participants are asked to label which one exhibits more positive emotion and which clip is more intense. The three categories of audio sources are presented on three separate pages. The participants are always comparing chroma reconstructions to chroma reconstructions, MFCC reconstructions to MFCC reconstructions or MIDI renderings to MIDI renderings. Subjects never compare a reconstruction to the original audio. For each round, we randomly select a clip to repeat as a means of verification. If a user labels the duplicated verification clip differently during the round with the original audio, their data is removed from the dataset.

4 Experiments and Results

Our first set of experiments investigates the emotional information retained in some of the most common acoustic features used in Music-IR, MFCCs and chromagrams. As described above, users listen to a pair of clips that was reconstructed from features (MFCC or chroma) and rate which is more positive and which has more emotional intensity. We seek to quantify how much information about musical emotion is retained in these acoustic features by how strongly emotion ratings of the reconstructions correlate with that of the originals. We first relate the user ratings to musical tempo and mode, and then we explore which features exhibit high variance with changes in tempo and mode or are invariant to altering these musical qualities.

4.1 Perceptual Evaluation of Acoustic Features

Running the task for three days, we collected a total of 3661 completed HITs, and accepted 1426 for an approval rating of 39%, which is similar to previous work annotating music data with MTurk [16–18]. The final dataset contains 17112 individual song pair annotations, distributed among 457 unique Turkers, with each Turker completing on average ~ 2.5 HITs. With a total of 160 pairs, this equates to ~ 35.65 ratings per pair. HITs are rejected for completing the task too quickly (less than 5 minutes), failing to label the repeated verification pairs the same for the original versions, and failing too many previous HITs.

[6] `http://music.ece.drexel.edu/research/emotion/invariance`

While repeated clips were presented for both reconstruction pairs and originals, requiring identical ratings on the reconstructions ultimately proved to be too stringent, due to the nature of the reconstructed clips. For the original clips we required the repeated pair to have the same ratings for both the higher valence and higher arousal clips, and reversed the A/B presentation of the clips to ensure Turk users were not just selecting song A or song B for every pair to speed through the task.

For each pair and for each audio type, we compute the percentage of subjects that rated clip A as more positive (valence) and the percentage that labeled clip A as more intense (arousal)

$$p_v = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\{A_n = \text{HigherValence}\}, p_a = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\{A_n = \text{HigherArousal}\} \quad (1)$$

where N is the total number of annotations for a given clip, p_v is the percentage of annotators that labeled clip A as higher valence, and p_a is the percentage of annotators that labeled clip A as higher arousal. For each song pair, we then compare the percentage of Turkers who rated song A as more positive in the original audio to those who rated song A more positive in the reconstructions, yielding the normalized difference error for all songs.

Audio Source	Normalized Difference Error	
	Valence	Arousal
MFCC Reconstructions	0.133 ± 0.094	$\mathbf{0.104 \pm 0.080}$
Chroma Reconstructions	$\mathbf{0.120 \pm 0.095}$	0.121 ± 0.082

Table 1. Normalized difference error between the valence/arousal ratings for the reconstructions versus the originals.

In Table 1, we show the error statistics for the deviation between the two groups. The paired ratings of each type are also verified with a paired Student's t-test to verify that they do not fall under the alternative hypothesis that there is a significant change, but as we are looking for proof that there is no change, average error remains the best indicator.

4.2 Relationships Between Muiscal Attributes and Emotional Affect

Next we analyze the data for trends relating major/minor modes and tempo to valence and arousal. In Section 2, we discussed the general trend of major tonality being associated with positive emotional affect and higher tempo corresponding to an increase in arousal or valence.

We divide our entire dataset S into a subset $M \subset S$ that consists of pairs that contain one major mode song and one minor mode song, as well as a subset $T \subset S$ in which pairs differ in tempo by more than 10 beats per minute (bpm). For subset M, we calculate what percentage of users labeled the major song as more positive and what percentage of users label the major song as more intense. For subset T, we similarly determine whether the faster song is more intense and whether the faster song is happier according to the users. Looking at Table 2, we conclude that the results are commensurate with the findings from the various psychology studies referenced in Section 2, namely that major songs are happier and faster songs are more intense.

Null Hypothesis	Agreement Ratio
Major Key Labeled as More Positive Valence	**0.667**
Faster Tempo Labeled More Positive Valence	**0.570**
Major Key Labeled as More Positive Arousal	0.528
Faster Tempo Labeled as More Positive Arousal	0.498

Table 2. Percentage of paired comparisons that yielded the desired perceptual result for mode and tempo.

One area where we expected larger agreement is the relationship between tempo and intensity. We only have the beats per minute for each song, and we label the faster song as the one with a higher bpm. The note lengths and emphasis in relation to the tempo are disregarded in this analysis and may be a source of uncertainty in the result. Depending upon the predominant note value (quarter/eighth/sixteenth), a slower tempo can sound faster than a song with a higher number of beats per minute. These are two different compositions, not the same clip at two different tempos.

4.3 Identifying Informative Feature Domains

When using features to understand certain perceptual qualities of music, it is important to know how those features relate to changes in the perceptual qualities being studied. We want to find appropriate variances and invariances as they relate to a perceptual quality. For example, if emotion is invariant to key, if the key changes, the features should also be invariant to that key change. We want correlation in variance as well. If the emotion of the audio changes, we want the features that describe it to change in conjunction with it. In order to investigate these variances and invariances, we use a feature set from prior work [3], as well as a set of features from the MIR-toolbox. Using the Beatles' clips, we generate changes in key, tempo, and mode to investigate possible corresponding differences in features. For key, the original was compared with transposed versions a 5th above and below. For tempo, the original was compared with versions at 75% and 133% of the original tempo. For mode, we shifted all the minor songs

to major and all the major songs to natural minor and compared the full dataset in major vs. the full dataset in minor.

Because the features contain different dimensions and have different ranges, looking at differences in their direct results does not allow for proper comparison between them. In order to draw proper comparisons, the features are normalized over dimension and range.

Given 2 feature vectors over time $F_1 \in \mathbb{R}^{M \times N}$ and $F_2 \in \mathbb{R}^{M \times N}$, we normalize the content over the vectors' shared range.

$$F_1' = \frac{F_1 - \min(F_1 \cup F_2)}{\max(F_1 \cup F_2)}, F_2' = \frac{F_2 - \min(F_1 \cup F_2)}{\max(F_1 \cup F_2)}, \tag{2}$$

The mean for each dimension is calculated, creating mean vectors $\mu_1 \in \mathbb{R}^{N \times 1}$ and $\mu_2 \in \mathbb{R}^{N \times 1}$. The average feature change across all dimensions is then computed.

$$FeatureChange = \frac{1}{N} \sum_{n=1}^{N} |\mu_1(n) - \mu_2(n)|, \tag{3}$$

If this $FeatureChange$ value is low, it means that the feature is invariant to the musical change being presented. In Table 3 we observe that features that exhibit higher variance to the specified change (tempo up/down, key up/down, and mode shift) should be more effective in computational models that are sensitive to these parameters. Several intuitive features including onsets, RMS energy, and beat spectrum emerge as the most variant features to tempo. Conversely, it is intuitive that features like mode and tonal center do not vary much with tempo.

Tempo Up		Tempo Down		Key Up		Key Down		Mode Shift	
Feature Domain	Feature Change	Feature Domain	Feature Change	Feature Domain	Feature Change	Feature Domain	Feature Change	Feature Domain	Feature Change
Onsets	0.127	Onsets	0.126	Key	0.142	Key	0.145	Mode	0.142
Beat Spec.	0.081	Beat Spec.	0.078	Beat Spec.	0.134	Beat Spec.	0.131	Tonal Cent.	0.114
RMS Energy	0.049	RMS	0.050	Tonal Cent.	0.105	Tonal Cent.	0.102	Beat Spec.	0.103
HCDF	0.024	HCDF	0.022	MFCC	0.084	MFCC	0.178	Key	0.063
xChroma	0.024	xChroma	0.021	Zerocross	0.081	Zerocross	0.064	Chroma	0.047
Roughness	0.023	Roughness	0.019	Chroma	0.055	Chroma	0.051	MFCC	0.030
Zerocross	0.022	SSD	0.017	Contrast	0.054	Contrast	0.049	Brightness	0.019
Brightness	0.021	MFCC	0.016	Regularity	0.050	xChroma	0.048	Onsets	0.015
SSD	0.021	Brightness	0.015	xChroma	0.038	Regularity	0.045	Attacktime	0.014
MFCC	0.017	Zerocross	0.015	Mode	0.038	SSD	0.041	Regularity	0.013
Chroma	0.014	Chroma	0.014	Brightness	0.037	Brightness	0.041	Zerocross	0.012
Key	0.013	Key	0.014	SSD	0.036	Mode	0.040	Contrast	0.011
S. Contrast	0.012	Regularity	0.011	Attacktime	0.030	Attacktime	0.026	xChroma	0.011
Regularity	0.012	Contrast	0.010	RMS	0.021	Roughness	0.023	SSD	0.010
Fluctuation	0.011	Fluctuation	0.009	Roughness	0.021	Onsets	0.020	RMS	0.009
Attacktime	0.010	Mode	0.007	Onsets	0.017	RMS	0.017	Attack Slope	0.008
Mode	0.009	Attacktime	0.007	Attack Slope	0.015	HCDF	0.015	Roughness	0.007
Tonal Cent.	0.007	Tonal Cent.	0.006	HCDF	0.012	Attack Slope	0.009	HCDF	0.006
Attack Slope	0.006	Attack Slope	0.005	Fluctuation	0.008	Fluctuation	0.008	Fluctuation	0.002

Table 3. Normalized feature change with respect to musical mode and tempo alterations.

5 Discussion and Future Work

In this paper, we have provided a perceptual evaluation of emotional content in audio reconstructions from acoustic features, and at the time of writing we know of no other work that has performed such experiments. In addition, we have related our findings to those of previous work showing correlation between major keys and increased positive emotion as well as increased tempo and increased positive emotion and activity. For tempo, mode and key we have provided a variational analysis for a large number of acoustic features. The findings we presented should be informative for future computational investigations in modeling emotions in music using content based methods.

References

1. Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull, "Music emotion recognition: A state of the art review," in *ISMIR*, Utrecht, Netherlands, 2010.
2. E. M. Schmidt and Y. E. Kim, "Modeling musical emotion dynamics with conditional random fields," in *ISMIR*, Miami, FL, 2011.
3. ——, "Prediction of time-varying musical mood distributions from audio," in *ISMIR*, Utrecht, Netherlands, 2010.
4. E. M. Schmidt, D. Turnbull, and Y. E. Kim, "Feature selection for content-based, time-varying musical emotion regression," in *ACM MIR*, Philadelphia, PA, 2010.
5. E. M. Schmidt and Y. E. Kim, "Learning emotion-based acoustic features with deep belief networks," in *WASPAA*, New Paltz, NY, 2011.
6. P. N. Juslin, J. Karlsson, E. Lindström, A. Friberg, and E. Schoonderwaldt, "Play it again with feeling: Computer feedback in musical communication of emotions," *Journal of Experimental Psychology: Applied*, vol. 12, no. 2, pp. 79–95, 2006.
7. K. Hevner, "Experimental studies of the elements of expression in music," *American Journal of Psychology*, no. 48, pp. 246–268, 1936.
8. M. G. Rigg, "Speed as a determiner of musical mood," *Journal of Experimental Psychology*, vol. 27, pp. 566–571, 1940.
9. G. D. Webster and C. G. Weir, "Emotional responses to music: Interactive effects of mode, texture, and tempo," *Motivation and Emotion*, vol. 29, pp. 19–39, 2005.
10. R. E. Thayer, *The Biopsychology of Mood and Arousal.* Oxford, U.K.: Oxford Univ. Press, 1989.
11. D. M. Randel, *The Harvard dictionary of music / edited by Don Michael Randel*, 4th ed. Belknap Press of Harvard University Press, Cambridge, MA :, 2003.
12. L. Gagnon and I. Peretz, "Mode and tempo relative contributions to happy-sad judgements in equitone melodies," *Cognition & Emotion*, vol. 17, no. 1, pp. 25–40, 2003.
13. S. Dalla Bella, I. Peretz, L. Rousseau, and N. Gosselin, "A developmental study of the affective value of tempo and mode in music." *Cognition*, vol. 80, no. 3, Jul. 2001.
14. G. Gerardi and L. Gerken, "The development of affective responses to modality and melodic contour," *Music Perception*, vol. 12, no. 3, pp. 279–290, 1995.
15. E. G. S. Gabriela Husain, WIlliam Thompson, "Effects of musical tempo and mode on arousal, mood, and spatial abilities," *Music Perception*, vol. 20, no. 2, pp. 151–171, 2002.

16. J. A. Speck, E. M. Schmidt, B. G. Morton, and Y. E. Kim, "A comparative study of collaborative vs. traditional annotation methods," in *ISMIR*, Miami, Florida, 2011.
17. J. H. Lee, "Crowdsourcing music similarity judgments using mechanical turk," in *ISMIR*, Utrecht, Netherlands, 2010.
18. M. I. Mandel, D. Eck, and Y. Bengio, "Learning tags that vary within a song," in *ISMIR*, Utrecht, Netherlands, 2010.

Oral session 5:
Music Information Retrieval

Automatic Identification of Samples in Hip Hop Music

Jan Van Balen[1], Martín Haro[2], and Joan Serrà[3] *

[1] Dept of Information and Computing Sciences, Utrecht University, the Netherlands
[2] Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
[3] Artificial Intelligence Research Institute (IIIA-CSIC), Bellaterra, Barcelona, Spain
j.m.h.vanbalen@uu.nl, martin.haro@upf.edu, jserra@iiia.csic.es

Abstract. Digital sampling can be defined as the use of a fragment of another artist's recording in a new work, and is common practice in popular music production since the 1980's. Knowledge on the origins of samples holds valuable musicological information, which could in turn be used to organise music collections. Yet the automatic recognition of samples has not been addressed in the music retrieval community. In this paper, we introduce the problem, situate it in the field of content-based music retrieval and present a first strategy. Evaluation confirms that our modified optimised fingerprinting approach is indeed a viable strategy.

Keywords: Digital Sampling, Sample Detection, Sample Identification, Sample Recognition, Content-based Music Retrieval

1 Introduction

Digital sampling, as a creative tool in composition and music production, can be defined as the use of a fragment of another artist's recording in a new work. The practice of digital sampling has been ongoing for well over two decades, and has become widespread amongst mainstream artists and genres, including hip hop, electronic, dance, pop, and rock [11]. Information on the origin of samples holds valuable insights in the inspirations and musical resources of an artist. Furthermore, such information could be used to enrich music collections, e.g. for music recommendation purposes. However, in the context of music processing and retrieval, the topic of automatic sample recognition seems to be largely unaddressed [5, 12].

The Oxford Music Dictionary defines sampling as "the process in which a sound is taken directly from a recorded medium and transposed onto a new recording" [8]. As a tool for composition, it first appeared when *musique concrète* artists of the 1950's started assembling tapes of previously released music recordings and radio broadcasts in musical collages. The phenomenon reappeared when

* This research was done between 1/2011 and 9/2011 at the Music Technology Group at Universitat Pompeu Fabra in Barcelona, Spain. The authors would like to thank Perfecto Herrera and Xavier Serra for their advice and support. JS acknowledges JAEDOC069/2010 from Consejo Superior de Investigaciones Científicas and 2009-SGR-1434 from Generalitat de Catalunya. MH acknowledges FP7-ICT-2011.1.5-287711.

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

DJ's in New York started using their vinyl players to repeat and mix parts of popular recordings to provide a continuous stream of music for the dancing crowd. The breakthrough of sampling followed the invention of the digital sampler around 1980, when producers started using it to isolate, manipulate, and combine portions of others' recordings to obtain entirely new sonic creations [6, 13]. The possibilities that the sampler brought to the studio have played a role in the appearance of several new genres in electronic music, including hip hop, house music in the late 90's (from which a large part of electronic dance music originates), jungle (a precursor of drum&bass music), dub, and trip hop.

1.1 Motivations for Research on Sampling

A first motivation to undertake the automatic recognition of samples originates in the belief that the musicological study of popular music would be incomplete without the study of samples and their origins. Sample recognition provides a direct insight into the inspirations and musical resources of an artist, and reveals some details about his or her composition methods and choices made in the production. Moreover, alongside recent advances in folk song [16] and version identification [14] research, it can be applied to trace musical ideas and observe musical re-use in the recorded history of the last two decades.

Samples also hold valuable information on the level of genres and communities, revealing cultural influences and dependence. Researchers have studied the way hip hop has often sampled 60's and 70's African-American artists [6] and, more recently, Bryan and Wang [2] analysed musical influence networks in sample-based music, inferred from a unique dataset provided by the WhoSampled web project. Such annotated collections exist indeed, but they are assembled through hours of manual introduction by amateur enthousiasts. It is clear that an automated approach could both widen and deepen the body of information on sample networks.

As the amount of accessible multimedia and the size of personal collections continue to grow, sample recognition from raw audio also provides a new way to bring structure in the organization of large music databases, complementing a great amount of existing research in this direction [5, 12]. Finally, sample recognition could serve legal purposes. Copyright considerations have always been an important motivation to understand sampling as a cultural phenomenon; a large part of the academic research on sampling is focused on copyright and law [11].

1.2 Requirements for a Sample Recognition System

Typically observed parameters controlling playback in samplers include filtering parameters, playback speed, and level envelope controls ('ADSR'). Filtering can be used by producers to maintain only the most interesting part of a sample. Playback speed may be changed to optimise the tempo (time-stretching), pitch (transposition), and/or mood of samples. Naturally, each of these operations complicates their automatic recognition. In addition, samples may be as short as one second or less, and do not necessarily contain tonal information. Moreover,

given that it is not unusual for two or more layers to appear at the same time in a mix, the energy of the added layers can be greater than that of the sample. This further complicates recognition. Overall, three important requirements for any sample recognition system should be: (1) The system is able to identify heavily manipulated query audio in a given music collection. This includes samples that are filtered, time-stretched, transposed, very short, tonal and non-tonal (i.e. purely percussive), processed with audio effects, and/or appear underneath a thick layer of other musical elements. (2) The system is able to perform this task for large collections. Finally, (3) the system is able to perform the task in a reasonable amount of time.

1.3 Scientific Background: Content-based Music Retrieval

Research in content-based music retrieval can be characterised according to *specificity* [5] and *granularity* [9]. Specificity refers to the degree of similarity between query and match. Tasks with a high specificity mean to retrieve almost identical documents, low specificity tasks look for vague matches that are similar with respect to some musical properties. Granularity refers to the difference between fragment-level and document-level retrieval. The problem of automatic sample recognition has a mid specificity and very low granularity (i.e. very short-time matches that are similar with respect to some musical properties). Given these characteristics, it relates to audio fingerprinting.

Audio fingerprinting systems attempt to identify unlabeled audio by matching a compact, content-based representation of it, the fingerprint, against a database of labeled fingerprints [3]. Just like fingerprinting systems, sample recognition systems should be designed to be robust to additive noise and several transformations. However, the deliberate transformations possible in sample-based music production, especially changes in pitch and tempo, suggest that the problem of sample recognition is in fact a less specific task.

Audio matching and version identification systems are typical mid specificity problems. Version identification systems assess if two musical recordings are different renditions of the same musical piece, usually taking changes in key, tempo and structure into account [14]. Audio matching works on a more granular level and includes remix recognition, amongst other tasks [4, 9]. Many of these systems use chroma features [5, 12]. These descriptions of the pitch content of audio are generally not invariant to the addition of other musical layers, and require the audio to be tonal. This is often not the case with samples. We therefore believe sample recognition should be cast as a new problem with unique requirements, for which the existing tools are not entirely suitable.

2 Experiments

2.1 Evaluation Methodology

We now present a first approach to the automatic identification of samples [15]. Given a query song in raw audio format, the experiments aim to retrieve a ranked list of candidate files with the sampled songs first.

To narrow down the experiments, only samples used in hip hop music were considered, as hip hop is the first and most famous genre to be built on samples [6] (though regarding sample origins, there were no genre restrictions). An evaluation music collection was established, consisting of 76 query tracks and 68 candidate tracks [15]. The set includes 104 sample relations (expert confirmed cases of sampling). Additionally, 320 'noise' files similar to the candidates in genre and length were added to challenge the system. Aiming at representativeness, the ground truth was chosen to include both short and long samples, tonal and percussive samples, and isolated samples (the only layer in the mix) as well as background samples. So-called 'interpolations', i.e. samples that have been re-recorded in the studio, were not included, nor were non-musical samples (e.g. film dialogue). This ground truth was composed using valuable information from specialized internet sites, especially WhoSampled[4] and Hip Hop is Read[5]. As the experiment's evaluation metric, the mean average precision (MAP) was chosen [10]. A random baseline of 0.017 was found over 100 iterations, with a standard deviation of 0.007.

2.2 Optimisation of a State-of-the-Art Audio Fingerprinting System

In a first experiment, a state-of-the-art fingerprinting system was chosen and optimised to perform our task. We chose to work with the spectral peak-based audio fingerprinting system designed by Wang [17]. A fingerprinting system was chosen because of the chroma argument in Section 1.3. The landmark-based system was chosen because of its robustness to noise and distortions and the alleged 'transparency' of the spectral peak-based representation (Table 1): Wang reports that, even with a large database, the system is able to correctly identify each of several tracks mixed together.

Table 1. Strengths and weaknesses of spectral peak-based fingerprints in the context of sample identification.

Strengths	Weaknesses
– High proven robustness to noise and distortions. – Ability to identify music from only a very short audio segment. – 'Transparent' fingerprints: ability to identify multiple fragments played at once. – Does not explicitly require tonal content.	– Not designed for transposed or time-stretched audio. – Designed to identify tonal content in a noisy context, fingerprinting drum samples requires the opposite. – Can percussive recordings be represented by just spectral peaks at all?

[4] http://www.whosampled.com/
[5] http://www.hiphopisread.com/

As in most other fingerprinting systems, the landmark-based system consists of an extraction and a matching component. Briefly summarized, the extraction component takes the short time Fourier transform (STFT) of audio segments and selects from the obtained spectrogram a uniform constellation of prominent spectral peaks. The time-frequency tuples with peak locations are paired in 4-dimensional 'landmarks', which are then indexed as a start time stored under a certain hash code for efficient lookup by the matching component. The matching component retrieves for all candidate files the landmarks that are identical to those extracted from the query. Query and candidate audio segments match if corresponding landmarks show consistent start times [17].

A Matlab implementation of this algorithm has been made available by Ellis[6]. It works by the same principles as [17], and features a range of parameters to control the implementation-level operation of the system. Important STFT parameters are the audio sample rate and the FFT size. The number of selected spectral peaks is governed by the desired density of peaks in the time domain and the peak spacing in the frequency domain. The number of resulting landmarks is governed by three parameters: the pairing horizons in the frequency and time domain, and the maximum number of formed pairs per spectral peak.

A wrapper function was written to slice the query audio into short fixed length chunks, overlapping with a hop size of one second, before feeding it to the fingerprinting system. A distance function is also required for evaluation using the MAP. Two distance functions are used, an absolute distance $d_a = \frac{1}{m+1}$, function of the number of matching landmarks m, and a normalized distance $d_n = \frac{l-m}{l}$, weighted by the number of extracted landmarks l.

Because of constraints in time and computational power, optimising the entire system in an extensive grid search would not be feasible. Rather, we have performed a large number of tests to optimise the most influential parameters. Table 2 summarizes the optimisation process, more details can be found in [15]. The resulting MAPs were 0.228 and 0.218, depending on the distance functions used (note that both are well beyond the random baseline mentioned before). Interestingly, better performance was achieved for lower sample rates. The optimal density of peaks and number of pairs per peak are also significantly larger than the default values, corresponding to many more extracted landmarks per second. This requires more computation time for both extraction and matching, and a requires for a higher number of extracted landmarks to be stored in the system's memory.

2.3 Constant Q Fingerprints

The MAP of around 0.22 is low for a retrieval task but promising as a first result. The system retrieves a correct best match for around 15 of the 76 queries. These matches include both percussive and tonal samples. However, due to the lowering of the sample rate, some resolution is lost. Not only does this discard valuable data, the total amount of information in the landmarks also goes down

[6] `http://labrosa.ee.columbia.edu/matlab/fingerprint/`

Table 2. Some of the intermediate results in the optimisation of the audio fingerprinting system by Wang as implemented by Ellis [15]. The first row shows default settings with its resulting performance.

pairs/pk	pk density (s^{-1})	pk spacing (bins)	sample rate (Hz)	FFT size (ms)	MAP_n (d_n)	MAP_a (d_a)
3	10	30	8,000	64	0.114	0.116
10	10	30	8,000	64	0.117	0.110
10	36	30	8,000	64	0.118	0.133
10	36	30	2,000	64	0.176	0.162
10	36	30	2,000	128	0.228	0.218

as the range of possible frequency values decreases. We now did a number of tests using a constant Q transform (CQT) [1] instead of a Fourier transform. We would like to consider all frequencies up to the default 8,000 Hz but make the lower frequencies more important, as they contributed more to the best performance so far. The constant Q representation, in which frequency bins are logarithmically spaced, allows us to do so. The CQT also suits the logarithmic representation of frequency in the human auditory system.

We used another Matlab script by Ellis[7] that implements a fast algorithm to compute the CQT and integrated it in the fingerprinting system. A brief optimisation of the new parameters returns an optimal MAP of 0.21 at a sample rate of 8,000 Hz. This is not an improvement in terms of the MAP, but loss of information in the landmark is now avoided (the amount of possible frequency values is restored), amending the system's scalability.

2.4 Repitching Fingerprints

In a last set of tests, a first attempt was made to deal with repitched samples. Artists often time-stretch and pitch-shift samples by changing their playback speed. As a result, the samples' pitch and tempo are changed by the same factor. Algorithms for independent pitch-shifting and time-stretching without audible artifacts have only been around for less than a decade, after phase coherence and transient processing problems were overcome. Even now, repitching is still popular practice amongst producers, as inspection of the ground truth music collection confirms. In parallel to our research [15], fingerprinting of pitch-shifted audio has been studied by Fenet et al. [7] in a comparable way, but the approach does not consider pitch shifts greater than 5%, and does not yet deal with any associated time-stretching.

The most straightforward method to deal with repitching is to repitch query audio several times and perform a search for each of the copies. Alternatively, the extracted landmarks themselves can also be repitched, through the appropriate scaling of time and frequency components (multiplying the time values

[7] See `http://www.ee.columbia.edu/~dpwe/resources/matlab/sgram/` and `http://labrosa.ee.columbia.edu/matlab/sgram/logfsgram.m`

Table 3. Results of experiments using repitching of both the query audio and its extracted landmarks to search for repitched samples.

N	ΔR (st)	r (st)	MAP_n	MAP_a
-	-	0	0.211	0.170
0	-	0.5	0.268	0.288
5	1.0	0.5	0.341	0.334
9	0.5	0.5	0.373	0.390

and dividing the frequency values, or vice versa). This way the extraction needs to be done only once. We have performed three tests in which both methods are combined: all query audio is resampled several times, to obtain N copies, all pitched ΔR semitones apart. For each copy of the query audio, landmarks are then extracted, duplicated and rescaled to include all possible landmarks repitched between $r = 0.5$ semitones up and down. This is feasible because of the finite resolution in time and frequency.

The results for repitching experiments are shown in Table 3. We have obtained a best performance of MAP_n equal to 0.390 for the experiment with $N = 9$ repitched queries, $\Delta R = 0.5$ semitones apart every query. This results in a total searched pitch range of 2.5 semitones up and down, or $\pm 15\%$. Noticeably, a MAP of 0.390 is low, yet it is in the range of some early version identification systems, or perhaps even better [14].

3 Discussion

To the best of our knowledge, this is the first research to address the problem of automatic sample identification. The problem has been defined and situated in the broader context of sampling as a musical phenomenon and the requirements that a sample identification system should meet have been listed. A state-of-the-art fingerprinting system has been adapted, optimised, and modified to address the task. Many challenges have to be dealt with and not all of them have been met, but the obtained performance of 0.39 is promising and unmistakably better than the precision obtained without taking repitching into account [15]. Overall, our approach is a substantial first step in the considered task.

Our system retrieved a correct best match for 29 of the 76 queries, amongst which 9 percussive samples and at least 8 repitched samples. A more detailed characterisation of the unrecognised samples is time-consuming but will make a very informative next step in future work. Furthermore, we suggest to perform tests with a more extensively annotated dataset, in order to assess what types of samples are most challenging to identify, and perhaps a larger number of ground truth relations. This will allow to relate performance and the established requirements more closely and lead to better results, paving the road for research such as reliable fingerprinting of percussive audio, sample recognition based on cognitive models, or the analysis of typical features of sampled audio.

References

1. Brown, J. C.: *Calculation of a Constant Q Spectral Transform*, The Journal of the Acoustical Society of America, vol. 89, no. 1, p. 425 (1991)
2. Bryan, N. J. and Wang, G.: *Musical Influence Network Analysis and Rank of Sample-Based Music*, in Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR), pp. 329-334 (2011)
3. Cano, P., Battle, E., Kalker, T. and Haitsma, J.: *A Review of Audio Fingerprinting* The Journal of VLSI Signal Processing-Systems for Signal, Image, and Video Technology, vol. 41, no. 3, pp. 271-284 (2005)
4. Casey, M. and Slaney, M.: *Fast Recognition of Remixed Music Audio*, in Acoustics Speech and Signal Processing 2007 ICASSP 2007 IEEE International Conference on, vol. 4, no. 12, pp. 300-1 (2007)
5. Casey, M., R. Veltkamp, Goto, M., Leman, M., Rhodes, C. and Slaney, M.: *Content-Based Music Information Retrieval: Current Directions and Future Challenges*, Proceedings of the IEEE, vol. 96, no. 4, pp. 668-696 (2008)
6. Demers, J.: *Sampling the 1970s in Hip-Hop*, Popular Music, vol. 22, no. 1, pp. 41-56 (2003)
7. Fenet, S., Richard, G., and Grenier Y.: *A Scalable Audio Fingerprint Method with Robustness to Pitch-Shifting*, in Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR), Miami, USA (2011)
8. Fulford-Jones, W.: *Sampling*, Grove Music Online. Oxford Music Online. `http://www.oxfordmusiconline.com/subscriber/article/grove/music/47228` (2011)
9. Grosche, P., Müller, M. and Serrà, J. *Audio Content-Based Music Retrieval*, in Multimodal Music Processing, Dagstuhl Publishing, Schloss Dagstuhl-Leibniz Zentrum für Informatik, Germany. Under Review.
10. Manning, C. D., Prabhakar, R. and Schutze, H.: *An Introduction to Information Retrieval*. Cambridge University Press, Cambridge (2008).
11. McKenna, T.: *Where Digital Music Technology and Law Collide - Contemporary Issues of Digital Sampling, Appropriation and Copyright Law*, Journal of Information Law and Technology, vol. 1, pp. 0-1 (2000)
12. Müller, M., Ellis, D., Klapuri, A. and Richard, G.: *Signal Processing for Music Analysis*, Selected Topics in Signal Processing, IEEE Journal of, vol. 0, no. 99, pp. 1–1 (2011)
13. Self, H.: *Digital Sampling: A Cultural Perspective*, UCLA Ent. L. Rev., vol. 9, p. 347 (2001)
14. Serrà, J., Gómez, E. and Herrera P.: *Audio Cover Song Identification and Similarity: Background, Approaches, Evaluation and Beyond*, in Advances in Music Information Retrieval, Springer, pp. 307–332 (2010)
15. Van Balen, J.: *Automatic Recognition of Samples in Musical Audio*. Master's thesis, Universitat Pompeu Fabra, Spain, `http://mtg.upf.edu/node/2342` (2011)
16. Wiering, F., Veltkamp, R.C., Garbers, J., Volk, A., Kranenburg, P. & Grijp, L.P.: *Modelling Folksong Melodies* Interdisciplinary Science Reviews, vol. 34, no. 2-3, pp. 154-171 (2009)
17. Wang, A.: *An Industrial Strength Audio Search Algorithm*, in Proceedings of the International Conference on Music Information Retrieval (ISMIR) (2003)

Novel use of the variogram for MFCCs modeling

Simone Sammartino, Lorenzo J. Tardón, and Isabel Barbancho

Dept. Ingeniería de Comunicaciones, E.T.S. Ingeniería de Telecomunicación,
Universidad de Málaga, Campus Universitario de Teatinos s/n, 29071, Málaga, Spain
{lorenzo,ibp}@ic.uma.es

Abstract. The paper describes two novel variants of the use of the variogram as summarizing tool for the MFCCs. A full variogram calculated on the second MFCC and a reduced variogram calculated on a subset of distance lags on the whole MFCCs matrix (first coefficient excluded), are proposed as tools to synthesize the timbre information of the MFCCs, for music similarity. Also, four different weighting functions are tested for the calculus of the (Euclidean) distance among the songs. The performance of the methods is evaluated by the application of the pseudo-objective evaluation of the MIREX AMS task, and compared with the scores of the methods submitted to the MIREX AMS 2011.

Keywords: Music similarity, MFCCs, variogram, MIREX AMS, musical genre classification

1 Introduction

The massive improvement of the Internet communication technology over the last years allowed the fast development of on-line games, multimedia playing and digital content sharing. The advances in the distribution of music contents led to the urgent need of a proper storage, labelling and indexation of the material, with the aim of an efficient access and retrieving of the items. One of the most demanded task is the automatic recommendation of music contents, aimed to help the user to choose a track with the highest degree of similarity with some defined references.

One of the fields where the MIR community is currently investing more resources are the so called *content-based* music recommendation systems where music similarity is evaluated on the basis of the calculus of a number of descriptors from time and frequency domain, and the derivation of some kind of feature patterns that are used as signature of the songs.

One of the most successfully used features to describe the spectral content of an audio signal are the Mel Frequency Cepstral Coefficients (MFCCs) [14]. These short-term spectral-based features are popularly employed to summarize the timbre content of the song and they are involved in most of the known algorithms for music similarity.

The MFCCs are calculated according to a recognized standard procedure: 1) calculus of the short-term spectrogram, 2) mapping of the spectrogram on

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

the Mel scale, through the application of a Mel frequencies bank filter, 3) transformation of the filtered spectrum to decibels and finally 4) compression of the resulting matrix by the application of the Discrete Cosine Transform.

Due to the own scheme of calculation of the MFCCs, the resulting descriptor is a matrix whose size depends both on the number of coefficients (fixed a priori) and the set of chunks in which the song has been fractioned during the windowing of the spectrogram. For this reason, the use of the MFCCs is usually associated with some kind of clustering of the coefficients, in order to represent the timbre descriptor as a fixed-size compressed matrix, to be employed directly as a standardized signature for the audio signals.

Logan and Salomon [10], employed the popular K-means method to cluster the MFCCs and used the means and covariance matrices of the centroids to define the song signature. Pampalk [13] proposed the use of the Gaussian Mixture Models (GMM) and the Expectation-Maximization (EM) approach, by modelling the probability distribution functions of the coefficients vectors. Aucouturier and Pachet [1] employed the Monte Carlo approach as clustering technique, Mandel and Ellis [11] used only one cluster from GMM, while Tzanetakis and Cook [17] simply extracted the mean and variance from each vector of Mel coefficients. In [15], Sammartino et al proposed the use of the variogram for MFCCs modelling. In this work, two novel variants of the calculus of the variogram are analyzed and their performance is evaluated with different setups.

The article is organized as follows: after this brief introduction on the music similarity framework, the use of the variogram as summarizing tool for the MFCCs is detailed in Section 2. The results of the evaluation of the methods proposed are presented in Section 3 and finally, some conclusions are drawn in Section 4.

2 The variogram

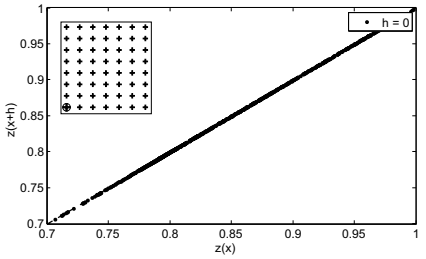
The variogram is a very popular tool in Geostatistics, widely employed to model the spatial continuity of environmental variables. Isaaks and Srivastava [5] affirm that "*Two data close to each other are more likely to have similar values than two data that are far apart.*" This characteristic is quantitatively defined as *spatial continuity*, referring to the spatial correlation of spatial variables.

2.1 The spatial variogram

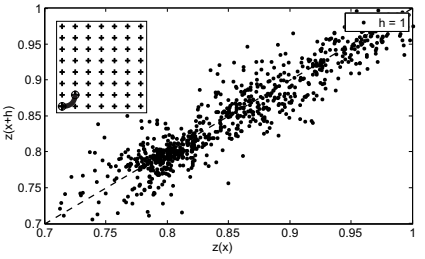
Let $z(x)$, with $x = 1, \ldots, n$ represents a set of n (regularly) sampled observations of a spatial phenomenon. The term x stands for the vector of spatial bi- or three-dimensional coordinates of the samples (generally unidimensional in the case of temporal variables).

One way to measure the spatial continuity of the samples is to observe how they behave when paired by their reciprocal distance. The h-scatter plot fulfills this target. It is the scatter plot of samples paired by a specific value of distance h [5].

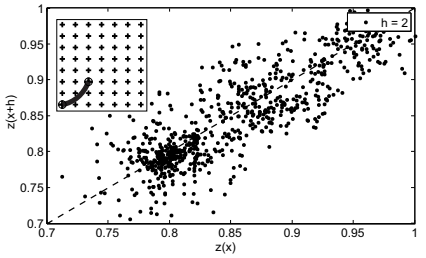
In Figure 1, three examples of h-scatter plots are shown. The samples are paired by a three distance vectors of $h = \{1, 2, 5\}$ (more precisely, it is $h = \{1, 2, 5\}$ in both x and y axis), and represented as points scattered over the bisector. Additionally, the h-scatter plot of the points coupled with themselves ($h = 0$) is presented (Fig. 1(a)). The bisector represents the geometrical locus of the pairs of samples separated by zero distance (all the samples paired with themselves).



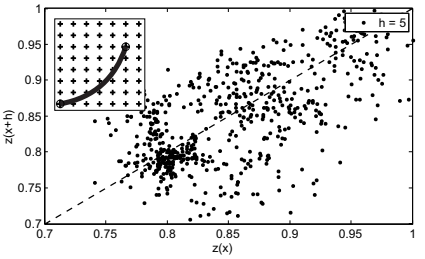
(a) The h-scatter plot calculated with samples paired with themselves ($h = 0$).



(b) The h-scatter plot calculated with samples paired at a distance $h = 1$.



(c) The h-scatter plot calculated with samples paired at a distance $h = 2$.



(d) The h-scatter plot calculated with samples paired at a distance $h = 5$.

Fig. 1. h-scatter plots of a set of regularly gridded spatial samples. The first plot is perfectly aligned with the diagonal. In fact the act of pairing samples at distance $h = 0$ means comparing each samples with itself. In the other plots the increase of the spread of the cloud is evident. In each plot, a small graph showing the pairing of the first sample is shown. The analysed data are a subset of the topographic data, provided by the US National Geophysical Data Centre (NOAA).

Following the definition of the spatial continuity, we can model how the spread of the clouds of points varies with the distance h. The greater the distance of the paired points, the fatter the cloud and the larger the difference between the samples of each pair. Therefore, it can be assumed that the spread of the cloud will range between zero, when the distance is null and the samples are paired with themselves, and a certain maximum extent, reached when the distance of the samples is large enough to fill the axes. At that point, even increasing the

distance, the spread of the cloud will not change substantially. The spread will achieve a steady value, with small oscillation around it. The paired samples will reach their reciprocal independence.

The way the spread of the cloud varies over the distance, resumes the law of spatial continuity of the samples analysed [5]. Three analogue approaches are employed in geostatistics, to model this law:

- The correlation coefficient of the pairs, whose variation with the distance is defined as the *correlogram*.
- The covariance and the corresponding *covariance function*.
- The moment of inertia and the corresponding *variogram*.

The general definition of the moment of inertia of two paired variables x and y, follows [5]:

$$T = \frac{1}{2n} \cdot \sum_{i=1}^{n} (x_i - y_i)^2 \tag{1}$$

where the factor $\frac{1}{2}$ refers to the perpendicular distance of the n samples to the diagonal.

Hence, the empirical variogram, or semivariance, of two paired variables ($z(x)$ and $z(x+h)$ separated by the distance h is defined as follows:

$$\gamma(h) = \frac{1}{2n(h)} \cdot \sum_{i=1}^{n(h)} (z(x_i) - z(x_i + h))^2 \tag{2}$$

where the number of pairs n is represented as a function of h, because their availability changes with the distance h. The term h is usually referred as the *lag*.

A typical variogram curve reflects the empirical assumption made for the h-scatter plot. It is zero at the origin, it increases with the lag distance and starts to flatten around a certain value of variance. In Figure 2, a typical empirical variogram is shown, together with the covariance function.

For interpolation purposes, the approximation of the law of spatial continuity for all the lags is often demanded. Then, the empirical variogram is usually asked to be fitted by some theoretical analytic models.

To infer the theoretical behavior of the experimental variogram, the samples of the spatial variable are considered as the realizations of a random function, a random variable, and a series of assumptions are drawn. In particular, the assumption on the stationarity of the random function is done. In conditions of a second order stationarity [18], the empirical variogram can be conveniently fitted by a family of functions (bounded authorized models), that allow to infer the information of the spatial continuity over the entire field. Two of the most popular models used for variogram fitting are the exponential and the spherical model [5].

A direct relation between the covariance function and the variogram can be defined. The covariance function starts at the variance of the random function

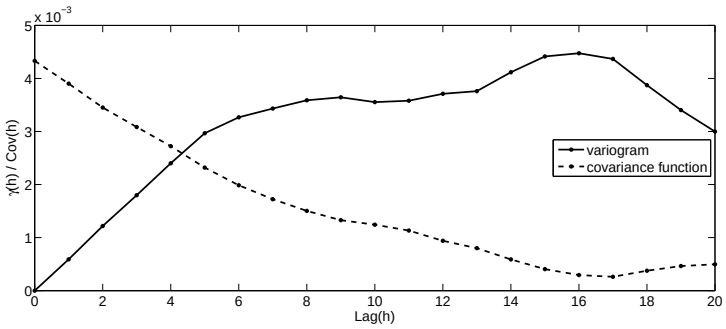


Fig. 2. A typical empirical variogram and its corresponding covariance function.

and decreases with the distance, tending to zero, when the samples of the random variable are sufficiently separated to be independent. Conversely, the variogram starts at zero, where the samples are at identical location and its variance is null, and increases with the distance, revealing the raise of the independence of the variable. It tends to the maximum degree of independence, that is the global variance of the random function [15].

The fit of the empirical variogram with the analytic models allows to parametrize the variogram function. Two main features are typically retrieved as descriptors of the shape of the theoretical variogram model: the *sill*, that is the variance at which the curve tends and the *range*, the lag value at which the sill is reached. A third very important parameter is the so called *nugget effect*. As seen, the theoretical value of the variogram at $h = 0$ is zero, because of the comparison of two different random variables at identical locations. However, in a practical experimental framework, a discontinuity of the empirical variogram at the short scale can be observed. This phenomenon is referred as the *small scale variability* [8]. The nugget effect is taken into account in the fit of the theoretical models by summing a certain quantity to the main model, such to shift the first lag to a level of variance higher than zero and cope with the small scale variability.

In Figure 3, a typical fitted theoretical model is shown, together with its main parameters.

2.2 The temporal variogram

Despite the variogram was born in a spatial statistics framework, it can be conveniently applied to time series data. Many authors [7, 4, 6] have dealt with the use of the variogram, coupled to classical signal processing techniques, as a tool for periodicity analysis of signals and time series analysis.

In the case of temporal signal processing, the distance parameter h is unidimensional and it represents the time lag among the samples. Unlike the spatial framework, where the samples are (regularly or not) distributed in the domain, in a temporal framework all the lag values are covered. The pairs availability is

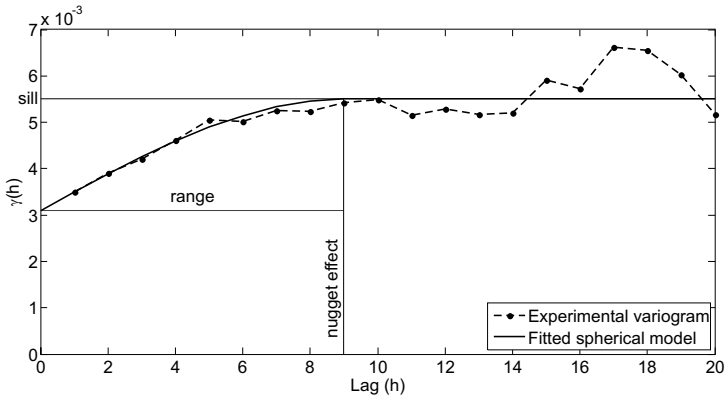


Fig. 3. An empirical variogram fitted by the analytical model, with the corresponding parameters.

a linear descending function with its maximum at lag $h = 1$, where the number of available pairs is $n - 1$ (where n represents the number of audio samples), ad its minimum, at lag $h = n - 1$, where the number of available pairs is 1.

For this reason, the reliability of the variogram values decreases with the lag. The variogram values estimated for the first lags are much more reliable than the last ones. Fortunately, the most revealing part of a variogram is indeed at the small scale, where it varies more, while a less interesting and rather constant behaviour is expressed at larger scales, just where the pairs availability decreases linearly and the estimation of this measure is less reliable.

In Figure 4, a typical temporal variogram is shown.

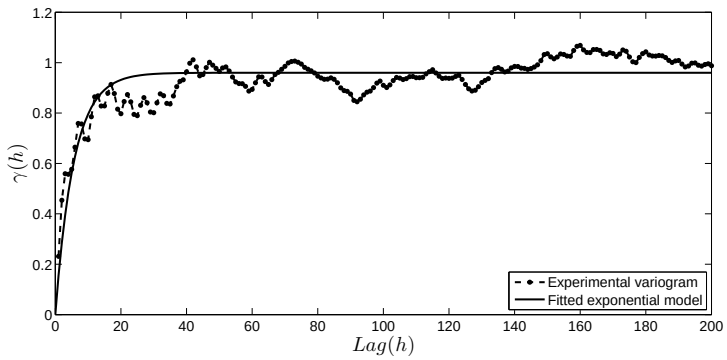


Fig. 4. A Typical temporal variogram. The experimental variogram is fitted by an exponential model.

Finally, when applied to audio signals, the variogram curve typically shows a periodical behaviour. In fact, the squared difference among the samples is affected by the periodicity of the signal itself and it is faithfully reflected by the variogram.

2.3 Variogram for MFCCs modeling

In this work, the temporal variogram is calculated on the MFCCs, as a tool for modelling the variation of the cepstral descriptor over the time fragments. A variant of the variogram proposed in [15] and a series of setups for the calculus of the distance are tested.

For the calculus of the MFCCs, the input signal is fractioned in a series of chunks with 1024 samples each, no windows overlap is employed and a hamming function is applied to each frame. The number of Mel filters (the triangular filterbank) is 40, while the number of DCT coefficients is 13. With such kind of configuration, one minute of audio signal corresponds to an MFCCs matrix of 13 x 2583 samples.

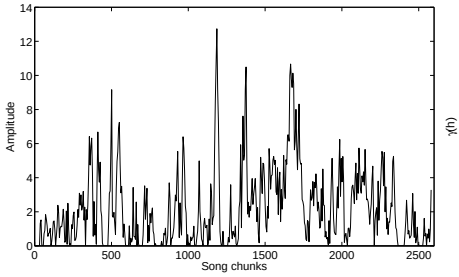
When the variogram is applied to the MFCCs, the lags values correspond to a temporal distance in terms of number of chunks in which the song has been fractioned. In order to achieve a standard measure to be employed in the quantitative comparison among the songs, each variogram is normalized by the global variance of the MFCC analysed. The result is an empirical variogram with an asymptotic tendency towards a reference variance of one. This is defined as *standardized variogram* [15].

The variogram is applied to the ISMIR 2004 Audio Description Contest (pre-MIREX) database for genre classification [2]: a set of about 700 songs, whose minimum and maximum duration was considered as 5 seconds and 5 minutes, respectively.

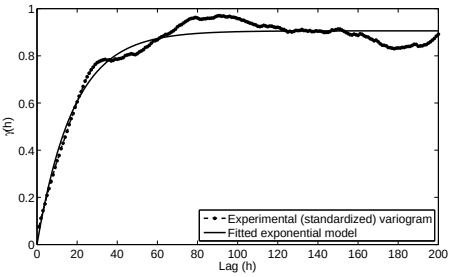
Full variogram The so called *full variogram* is the variogram of the second MFCC, calculated from lag 1 to 200. That is from the temporal pairwise distance corresponding to 1 chunk (1024 samples, about 23 ms) to the one corresponding to 200 chunks, that is about 4.6 seconds. The resulting unidimensional vector of 200 elements stands for the song signature. This approach implies a dimensionality reduction rate of about 93% (from about 2800 samples of the original MFCCs matrix (with size 215 x 13) to 200 samples of the variogram vector) in the case of the shortest audio fragment (5 seconds), and about 99.8% (from about 168000 samples of the original MFCCs matrix (with size 12919 x 13) to 200 samples of the variogram vector) in the case of the largest audio fragment with a maximum duration of 5 minutes.

In Figure 5, two examples of full variogram calculated on the second MFCC of two songs from the genre classical and electronic, are shown.

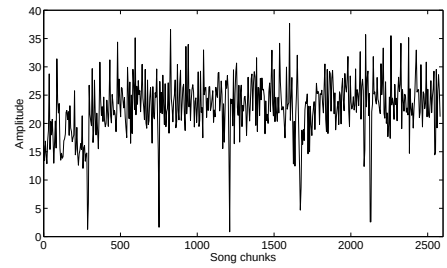
The large discrepancy expected by the comparison of two songs belonging to two very different genres, is reflected by the variogram analysis. The second MFCCs of the two songs are rather different: the one of the classical piece shows a



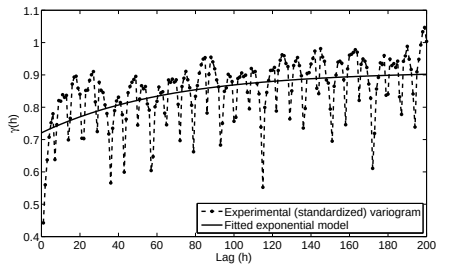
(a) The second MFCC of a classical piece.



(b) The standardized variogram of the second MFCC of a classical piece.



(c) The second MFCC of an electronic piece.



(d) The standardized variogram of the second MFCC of a electronic piece.

Fig. 5. Two examples of calculation of the full standardized variogram on the second MFCC of two songs, respectively from classical and electronic genre. The excerpts analyzed have a duration of 1 minute.

more structured and smoother variability, with few high frequency components and a hidden (or missing) periodicity, while the one of the electronic piece is much more fuzzy, with a large contribution of a high frequency variability and a marked periodical behaviour.

The corresponding variograms reflects very well the behaviour highlighted. The variogram of the classical piece reveals a very structured variability, with a high pairwise continuity at the small scale (the nugget effect is null) and a smoothly increasing variance with a clear asymptotic trend towards the range. Conversely, the variogram of the electronic piece is much more unstructured, with continuous periodic oscillation coupled to a very weak asymptotic trend. Its nugget effect is rather high.

Reduced variogram The *reduced variogram* is calculated on 12 MFCCs (from the second MFCC to the last one), on a reduced bunch of lags. A total amount of 20 lags are sampled with a logarithmically varying density, from 1 to 200, with the aim to concentrate the lags at the smallest scale, where most of the variance is expressed (see Figure 6).

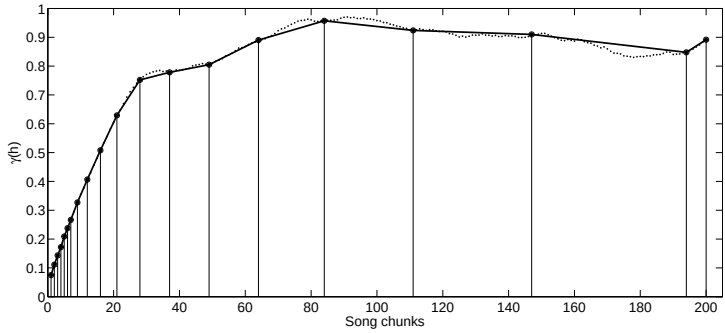


Fig. 6. The variogram for the classical piece of Figure 5(b), reduced by the lag sampling (thick line). Note the logarithmic distribution of the sampled lags.

The signature matrix is of size 12 x 20, resulting in a total amount of 240 elements (if stacked). The dimensionality reduction rate is quite the same of the full variogram. In Figure 7, the reduced versions of the full variogram of Figure 5 are shown.

The conclusions drawn for the reduced matrix of variograms are the same as for the full variogram. The classical piece shows smoother variograms, revealing a more structured variability and a high small scale pairwise continuity. Conversely, the electronic track reveals a more fuzzy variability structure and a marked periodical behaviour. In both cases, the reduction of the number of the lags keeps guaranteeing a faithful representation of the original full variogram.

2.4 Distance measurement

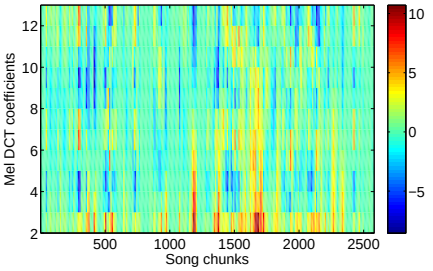
In order to estimate the degree of similarity of the songs, the signatures have to be numerically compared. In this work, a weighted Euclidean distance is used.

In general, the distance is calculated as follows:

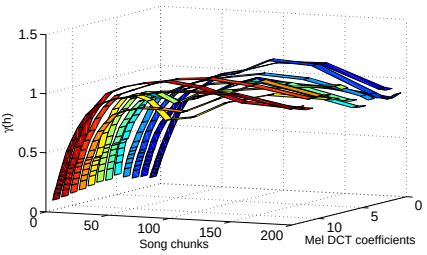
$$D_{i,j} = \sqrt{\sum_{k=1}^{n} \left((V_i(k) - V_j(k)) \cdot \omega(k) \right)^2} \tag{3}$$

where $V_i(k)$ and $V_j(k)$ are the values of the k-th lag of the variograms of two songs i and j, and $\omega(k)$ is the weight of the k-th lag, with a maximum number of lags n equal to 240, for a bi-dimensional reduced variogram and 200, for a full unidimensional variogram. Note that the bi-dimensional variogram is stacked into a unidimensional vector to simplify the calculus.

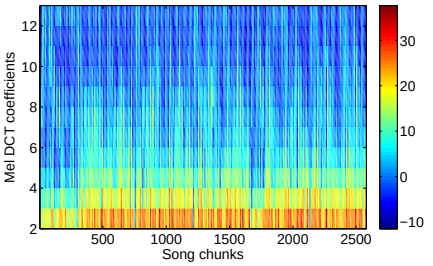
Actually, the variogram shows a maximum of information (in term of quality and reliableness) at the small scale. The most predominant meaning of the measure arises from the first lags, up to the achievement of the range, beyond which the variogram looses significance. For this reason, three different sets of weights



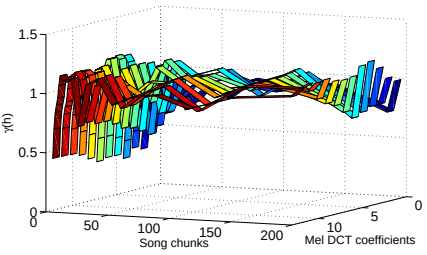
(a) The whole MFCCs matrix of a classical piece.



(b) The matrix of standardized variograms of the whole MFCCs matrix, reduced by the lags sampling. Classical piece.



(c) The whole MFCCs matrix of a electronic piece.



(d) The matrix of standardized variograms of the whole MFCCs matrix, reduced by the lags sampling. Electronic piece.

Fig. 7. Two examples of calculation of the matrix of standardized variograms of the whole MFCCs matrix of two songs, respectively from classical and electronic genre. The excerpts analyzed have a duration of 1 minute.

are proposed: a set of exponentially decreasing weights, a set of logarithmically decreasing weights and, finally, a set of linearly decreasing weights. A fourth unweighted variant of the distance is included.

In Figure 8, the three sets of weights are compared. Note that the vectors of weights represented here correspond to one of the stacked vectors of weights employed for the reduced variogram (20 lags).

Note that in any of the three cases, the weights are normalized such that their sum is 1.

3 Evaluation of the performance of the algorithms

The evaluation of the performance of the two variants of variogram, each with the four weighting functions, is implemented on the basis of the genre classification music database of the ISMIR 2004 Audio Description Contest [2].

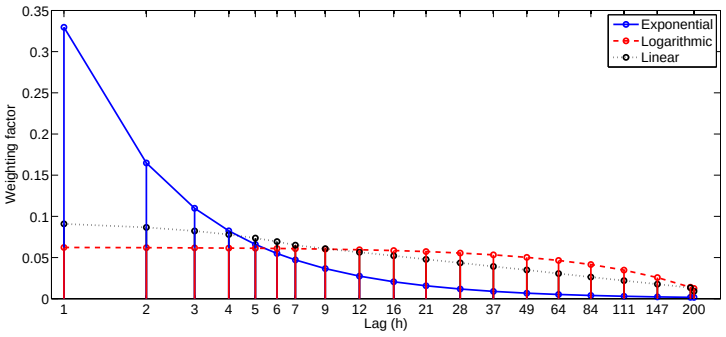


Fig. 8. The three vectors of weights employed for the calculus of the distance. Note that the shape of the linear weights is deformed by the scaling of the lag axis that is logarithmic.

The pseudo-objective evaluation [3], currently employed in the MIREX music similarity tasks, is performed. The matching rates of artist, album and (artist-filtered) genre, for the first 5, 10, 20 and 50 songs are calculated.

After sorting the list of songs according to the degree of similarity to the seed item (one of the songs of the collection, selected recursively), the pseudo-objective statistics are calculated as percentages of the songs of the list sharing the same artist, album or (artist-filtered) genre. These percentages are calculated four times, on a reference total of the first 5, 10, 20 and 50 songs of the list.

In order to compensate for the unequal distribution of items per category (artist, album or genre), the reference total is defined as the maximum between the defined reference (5, 10, 20 or 50) and the maximum number of available songs per category. For instance, if only 8 songs are available for a certain artist, the reference total for the calculus of the artist-based statistic has to be 5, for the first 5 songs, but it must be reduced to 8 for each of the higher counts (10, 20 or 50). In fact, the statistics would be negatively affected by considering the reference total as some values higher than the maximum allowed by the database itself. If the algorithm is able to return all the 8 correct correspondences for the artist into the first 8 positions, it has to be considered as best performing for each of the totals: 5/5 (for the first 5 items) and 8/8 (for the first 10, 20 or 50 items).

This procedure is recursively applied to all the songs of the collection, setting each time one of them as the seed song. Finally, the global score is calculated as the averaged mean of the scores obtained for each seed song. In Table 1, the matching scores for the two variants of variogram are shown.

The performance returned by the reduced variogram is globally better than the one of the full variogram, for any kind of weighting configuration. On the one hand, it is true that the full variogram returns a more complete information of the second (and most representative) MFCC, with respect to the reduced

	Full variogram				Reduced variogram			
	Exponential weights							
Artist	7.24	9.12	13.01	23.63	16.99	18.47	23.11	34.26
Album	5.29	8.52	14.23	26.20	13.25	19.14	26.30	37.69
Genre	**43.62**	**42.46**	**41.85**	**40.16**	**46.43**	**45.29**	**44.17**	**43.23**
	Logarithmic weights							
Artist	5.20	5.94	9.37	18.05	15.81	16.70	21.88	32.86
Album	3.86	5.70	10.43	19.40	12.62	17.83	25.72	37.03
Genre	**38.72**	**38.48**	**38.41**	**37.65**	**48.07**	**46.66**	**45.51**	**42.84**
	Linear weights							
Artist	5.14	5.81	8.90	18.45	16.51	17.70	22.40	33.77
Album	3.79	5.47	9.64	19.44	12.93	19.15	26.74	38.68
Genre	**39.25**	**38.44**	**38.40**	**37.79**	**47.84**	**46.94**	**44.81**	**42.73**
	No weights							
Artist	4.50	5.47	8.40	16.43	15.23	16.32	21.59	32.38
Album	3.42	5.65	9.23	17.97	13.03	17.55	25.23	36.07
Genre	**38.19**	**38.08**	**37.63**	**37.25**	**48.48**	**47.68**	**46.18**	**43.70**
	First 5	**First 10**	**First 20**	**First 50**	**First 5**	**First 10**	**First 20**	**First 50**

Table 1. Pseudo-objective statistics for the two variants of variogram calculation. Note that the genre scores are calculated on the artist-filtered subset. The genre results are in bold because these results are the ones that can be compared with the MIREX AMS 2011 results presented in table 2. It can be observed that the proposed methods perform quite well.

variant, that is calculated on a smaller bunch of lags. On the other hand, the completeness of the information based on the involvement of the complete set of MFCCs returns a more accurate description of the song analysed. Apparently, the loss of information due to the reduction of the lags is compensated by the gain derived by the employment of the complete MFCCs matrix.

Also, an inverse trend of variation of the scores is observed for the three different categories: artist, album and genre. In particular, the artist and album-based scores increase with the number of items considered, while for the genre the tendency is inverse. It basically depends on the availability of items per category. In fact, the probability of returning one song of the first 5, 10, 20 or 50 with the same genre of the seed song, is much higher than the one related to the other two categories. Actually, the genre-based statistic reflects the higher concentration power of the genre, that finds similar songs more easily, yet from the very first few items considered. Conversely, as finding the correct songs with the same artist or album is much harder, the larger the number of items considered, the higher the score obtained for these categories.

The best results for the full variogram (e.g.: 43.62%, obtained for the genre coincidence of the first 5 items of the list) have been obtained with the set of exponentially decreasing weights, where the first lags contribution is much higher than the others. Surprisingly, the result shown by the reduced variogram

is different: the best scores are referred to the null weighting of the distance, although the trend is not as clear as the case of the full variogram.

Table 2 shows the results of the pseudo-objective evaluation of the algorithms proposed to the Audio Music Similarity contest of the MIREX 2011 (only the artist-filtered genre scores) [12]. It can be observed that the scores obtained for the variogram-based approaches are in line with the reference represented by the MIREX Audio Music Similarity task. Although a direct quantitative comparison cannot be provided because of the differences in the test database used in the two frameworks, the variogram seems to return a rather reliable accuracy in the estimation of music similarity.

Method	First 5	First 10	First 20	First 50
STBD1	24.19	23.34	22.14	20.57
STBD2	23.55	22.56	21.61	19.98
STBD3	23.07	22.55	21.78	20.47
DM2	46.02	44.14	42.22	39.28
DM3	46.08	44.20	42.33	39.37
GKC1	23.45	22.55	21.57	20.01
HKHLL1	34.91	33.81	32.72	31.39
ML1	41.77	39.86	38.09	35.53
ML2	40.19	38.45	36.28	33.62
ML3	41.06	38.99	36.80	33.85
PS1	54.11	52.17	50.13	46.74
SSKS3	54.65	53.15	51.52	48.98
SSPK2	54.24	52.75	51.19	48.56
YL1	37.40	35.43	33.01	29.54

Table 2. Average artist-filtered genre scores of the algorithms proposed to the MIREX 2011 contest. The method acronyms correspond to the standard coding employed in the MIREX contest [12].

4 Conclusions and future works

In this paper, the use of the temporal variogram has been proposed as a tool to model the temporal variability of the Mel Frequency Cepstral Coefficients and it has been exploited to estimate music similarity.

After a brief description of the theory of the variogram analysis and its adaptation to a temporal framework, two different variants of the calculus of the variogram and four weighting functions for the calculus of the distance between the song signatures, have been proposed. Both the two variogram-based approaches have been tested on a reference database of songs, divided into six different genres. A pseudo-objective analysis has been computed in order to achieve a quantitative evaluation of the performance of the methods and propose a dis-

cussion on the results. Also, a comparison with the actual reference in term of algorithms aimed to perform music similarity, has been provided.

The reduced variogram returns better scores than the full variant, due to the more complete information given by the whole MFCCs matrix (the first MFCC excluded). This method seems not be really influenced by the kind of weighting function used for the calculus of the distance. The results are in line with the references of the MIREX 2011.

All the variograms analyzed so far are the results of the empirical calculus of the equation (2). In future development of the variogram-based approach, the automatic fitting of the theoretical models can be employed to try to resume the variogram function as a series of parameters. In particular, the nugget effect, the range and the sill of the theoretical models could be employed as low-level descriptors for classification purposes.

In order to test this concept, a very simple approximation has been carried out. A simple least square fit of the exponential model has been implemented to the variograms of the songs of the collection tested in the article, in order to achieve a first estimation of the nugget effect. Afterwards, it has been employed as a low-level descriptor, together with other popular MIR descriptors [16], and tested in a music genre classifier. The classifier employed was a simply knn-classifier, with $k = 5$ neighbors.

The results are rather encouraging. The performance of the nugget effect, although it has been estimated by a simply automatic fit of the experimental variograms, are in line with the one of other more popular features. The nugget effect reflects even a better behavior in some specific cases (as the example of the genre world).

As known, the automatic fitting of the empirical variograms is an actual matter of discussion and the issue is far from being resolved [9]. These preliminary results presented encourage to focus on the automatic fitting of the variogram models in order to obtain more robust descriptors to be conveniently used for MIR tasks.

Acknowledgments. This work has been funded by the Ministerio de Economía y Competitividad of the Spanish Government under Project TIN2010-21089-C03-02 and by the Ministerio de Industria, Turismo y Comercio of the Spanish Government under Project TSI-090100-2011-25.

References

1. Aucouturier, J.J., F., P.: Improving timbre similarity: How high is the sky? Journal of Negative Results in Speech and Audio Sciences 1(1) (2004)
2. Cano, P., Gmez, E., Gouyon, F., Herrera, P., Koppenberger, M., Ong, B., Serra, X., Streich, S., Wack, N.: Ismir 2004 audio description contest (2006), `http://mtg.upf.edu//files/publications/MTG-TR-2006-02.pdf`
3. Downie, S.J.: The music information retrieval evaluation exchange (mirex). D-Lib Magazine 12(12) (2006)

4. Haslett, J.: On the sample variogram and the sample autocovariance for non-stationary time series. The Statistician 46(4), 475–485 (1997)
5. Isaaks, E.H., Srivastava, M.R.: An Introduction to Applied Geostatistics. Oxford University Press, USA (January 1990)
6. Kacha, A., Grenez, F., Schoentgen, J., Benmahammed, K.: Dysphonic speech analysis using generalized variogram. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. (ICASSP '05). vol. 1, pp. 917–920 (2005)
7. Khachatryan, D., Bisgaard, S.: Some results on the variogram in time series analysis. Quality and Reliability Engineering International (March 2009)
8. Krige, D.G.: A statistical approach to some basic mine valuation problems on the witwatersrand. Journal of the Chemical, Metallurgical and Mining Society of South Africa 52(6), 119–139 (December 1951)
9. Li, S., Lu, W.: Automatic fit of the variogram. In: Third International Conference on Information and Computing (ICIC), 2010. vol. 4, pp. 129 –132 (june 2010)
10. Logan, B., Salomon, A.: A music similarity function based on signal analysis. In: Multimedia and Expo, 2001. ICME 2001. IEEE International Conference on. pp. 745 – 748 (aug 2001)
11. Mandel, M.I., Ellis, D.P.W.: Song-Level Features and Support Vector Machines for Music Classification. In: Reiss, J.D., Wiggins, G.A. (eds.) Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR). pp. 594–599 (Sep 2005)
12. MIREX2011: Audio music similarity and retrieval results (2011), http://www.music-ir.org/mirex/wiki/2011:Audio_Music_Similarity_and_Retrieval_Results
13. Pampalk, E.: Computational Models of Music Similarity and their Application to Music Information Retrieval. Ph.D. thesis, Vienna University of Technology, Vienna (March 2006)
14. Rabiner, L., Juang, B.H.: Fundamentals of speech recognition. Prentice-Hall, Inc., Upper Saddle River, NJ, USA (1993)
15. Sammartino, S., Tardón, L.J., de la Bandera, C., Barbancho, I., Barbancho, A.M.: The standardized variogram as a novel tool for music similarity evaluation. In: Proc. of Int. Symposium on Music Information Retrieval (ISMIR 2010). pp. 559–564 (2010)
16. Tardón, L.J., Sammartino, S., Barbancho, I.: Design of an efficient music-speech discriminator. Journal of the Acoustical Society of America 127(1) (2010)
17. Tzanetakis, G., Cook, P.: Musical genre classification of audio signals. Speech and Audio Processing, IEEE Transactions on 10(5), 293 – 302 (jul 2002)
18. Wackernagel, H.: Multivariate Geostatistics: An Introduction With Applications. Springer-Verlag Telos (January 1999)

Automatic String Detection for Bass Guitar and Electric Guitar

Jakob Abeßer

Fraunhofer IDMT,
Ehrenbergstr. 17, 98693 Ilmenau, Germany
{abr@idmt.fhg.de}
http://www.idmt.fraunhofer.de

Abstract. In this paper, we present a feature-based approach to automatically estimate the string number in recordings of the bass guitar and the electric guitar. We perform different experiments to evaluate the classification performance on isolated note recordings. First, we analyze how factors such as the instrument, the playing style, and the pick-up settings affect the performance of the classification system. Second, we investigate, how the classification performance can be improved by rejecting implausible classifications as well as aggregating the classification results over multiple adjacent time frames. The best results we obtained are f-measure values of $F = .93$ for the bass guitar (4 classes) and $F = .90$ for the electric guitar (6 classes).

Keywords: string classification, fretboard position, fingering, bass guitar, electric guitar, inharmonicity coefficient

1 Introduction

On string instruments such as the bass guitar or the guitar, most notes within the instrument's pitch range can be played at multiple positions on the instrument fretboard. Each *fretboard position* is defined by a unique string number and a fret number. Written music representations such as *common music notation* do not provide any information about the fretboard position where each note is to be played. Instead, musicians often have to choose an appropriate fretboard position based on their musical experience and stylistic preferences. The *tablature* representation, on the other hand, is specialized on the geometry of fretted string instruments such as the guitar or the bass guitar. It specifies the fretboard position for each note and thus resolves the ambiguity between note pitch and fretboard position. Fig. 1 illustrates a bass-line represented both as score and as tablature.

Conventional automatic music transcription algorithms extract score-related parameters such as the pitch, the onset, and the duration of each note. In order to analyze recordings of string instruments, the fretboard position needs to be estimated as an additional parameter. The ability to automatically estimate the fretboard position allows to generate a tablature and is therefore very useful for

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

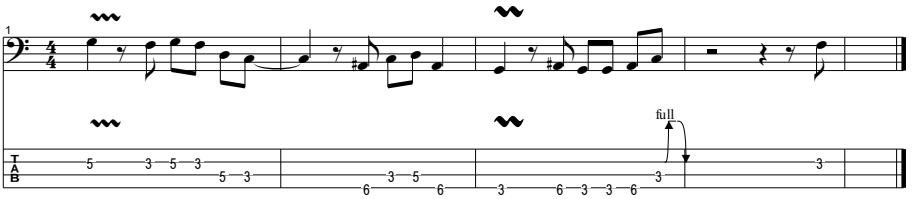


Fig. 1: Score and tablature representation of a bass-line. The four horizontal lines in the tablature correspond to the four strings with the tuning E1, A2, D2, and G2 (from bottom to top). The numbers correspond to the fret numbers on the strings that are to be played.

music assistance and music education software. This holds true especially if this software is used by beginners who are not familiar with reading musical scores. As will be discussed in Sect. 3, various methods for estimating the fretboard position were proposed in the literature so far, ranging from audio-based methods to methods that exploit the visual modality or that use attached sensors on the instrument. However, the exclusive focus on audio analysis methods for this purpose has several advantages: In music performance scenarios involving a bass guitar or electric guitar, the instrument signal is accessible since these instruments need to be amplified. In contrast, video recordings of performing musicians and the instrument neck are often limited in quality due to movement, shading, and varying lighting conditions on stage. Additional sensors that need to be attached to the instrument are often obtrusive to the musicians and affect their performance. Therefore, this paper focuses on a sole audio-based analysis.

This paper is structured as follows: We outline the goals and challenges of this work in Sect. 2. In Sect. 3, we discuss existing methods for estimating the fretboard position from string instrument recordings. A new approach solely based on audio-analysis is detailed in Sect. 4, starting with the spectral modeling of recorded bass and guitar notes in Sect. 4.1 and the note detection in Sect. 4.2. Based on the audio features explained in Sect. 4.2, we illustrate how the fretboard position is automatically estimated in Sect. 4.3. In Sect. 5, we present several evaluation experiments and discuss the obtained results. Finally, we conclude our work in Sect. 6.

2 Goals & Challenges

We aim to estimate the string number n_s from recorded notes of the bass guitar and the electric. Based on the note pitch P and the string number, we can apply knowledge on the instrument tuning to derive the fret number n_f and thus a complete description of the fretboard position. In the evaluation experiments described in Sect. 5, we investigate how the classification results are affected by separating the training and test data according to different criteria such as

the instruments, the pick-up (PU) settings, and the applied playing techniques. Furthermore, we analyze if a majority voting scheme that combines multiple string classification results for each note can improve the classification performance. The main challenge is to identify suitable audio features that allow to discriminate between notes that, on the one hand, have the same fundamental frequency f_0 but, on the other hand, are played on different strings. The automatic classification of the played string is difficult since the change of fingering alters the sonic properties of the recorded music signal only subtly. Classic non-parametric spectral estimation techniques such as the Short-Time Fourier Transform (STFT) are affected by the *spectral leakage* effect: the Fourier Transform of the applied window function limits the achievable frequency resolution to resolve closely located spectral peaks. In order to achieve a sufficiently high frequency resolution for estimating the harmonic frequencies of a note, rather larger time frames are necessary. The decreased time resolution is disadvantageous if notes are played with frequency modulation techniques such as bending or vibrato, which cause short-term fluctuations of the harmonic frequencies [1]. This problem is especially impeding in lower frequency bands. Thus, a system based on classic spectral estimation techniques is limited to analyze notes with only a slow-varying pitch, which can be a severe limitation for a real-word system. Since we focus on the bass guitar and the electric guitar, frequencies between 41.2 Hz and 659.3 Hz need to be investigated as potential f_0-candidates[1].

3 Related Work

In this section, we discuss previous work on the estimation of the played string and the fretboard position from bass and guitar recordings. First, we review methods that solely focus on analyzing the audio signal. Special focus is put on the phenomenon of inharmonicity. Then, we compare different hybrid methods that incorporate computer vision techniques, instrument enhancements, and sensors.

3.1 Audio Analysis

Penttinen et al. estimated the plucking point on a string by analyzing the delay times of the two waves on the string, which travel in opposite directions after the string is plucked [21]. This approach solely focuses on a time-domain analysis and is limited towards monophonic signals. In [3], Barbancho et al. presented an algorithm to estimate the string number from isolated guitar note recordings. The instrument samples used for evaluation were recorded using different playing techniques, different dynamic levels, and guitars with different string material. After the signal envelope is detected in the time-domain, spectral analysis based on STFT is applied to extract the spectral peaks. Then, various audio features

[1] This corresponds to the most commonly used bass guitar string tunings E2 to G3 and electric guitar string tuning E3 to E5, respectively, and a fret range up to the 12th fret position.

related to the timbre of the notes are extracted such as the spectral centroid, the relative harmonic amplitudes of the first four harmonics, and the inharmonicity coefficient (compare Sect. 3.1). Furthermore, the temporal evolution of the partial amplitudes is captured by fitting an exponentially decaying envelope function. Consequently, only one feature vector can be extracted for each note. As will be shown in Sect. 4.2, the presented approach in this paper allows to extract one feature vectors on a frame-level. This allows to accumulate classification results from multiple (adjacent) frames of the same note recording to improve the classification performance (compare Sect. 4.3). The authors of [3] reported diverse results from the classification experiments. However, they did not provide an overall performance measure to compare against. The performance of the applied classification algorithm strongly varied for different note pitch values as well as for different compilations of the training set in their experiments.

In [2], Barbancho et al. presented a system for polyphonic transcription of guitar chords, which also allows to estimate the fingering of the chord on the guitar. The authors investigated 330 different fingering configuration for the most common three-voiced and four-voiced guitar chords. A Hidden Markov Model (HMM) is used to model all fingering configurations as individual hidden states. Based on an existing multi-pitch estimation algorithm, harmonic saliency values are computed for all possible pitch values within the pitch range of the guitar. Then, these saliency values are used as observations for the HMM. The transitions between different hidden states are furthermore constrained by two models—a musicological model, which captures the likelihood of different chord changes, and an acoustic model, which measures the physical difficulty of changing the chord fingerings. The authors emphasized that the presented algorithm is limited towards the analysis of solo guitar recordings. However, it clearly outperformed a state-of-the-art chord transcription system. The applied dataset contained instrument samples of electric guitar and acoustic guitar. Maezawa et al. proposed a system for automatic string detection from isolated bowed violin note recordings in [16]. Similar to the bass guitar, the violin has 4 different strings, but in a higher pitch range. The authors analyzed monophonic violin recordings of various classical pieces with given score information. First, the audio signal is temporally aligned to the musical score. For the string classification, filterbank energies are used as audio features and a Gaussian Mixture Model (GMM) as classifier. The authors proposed two additional steps to increase the robustness of the classification. First, feature averaging and feature normalization are used. Then, a context-dependent error correction is applied, which is based on empirically observed rules how musicians choose the string number. The authors investigated how training and test with the same and different instruments and string types affect the classification scores (similar to Sect. 5). The highest F-measure value that was achieved for the string classification with 4 classes is $F = .86$.

Inharmonicity For musical instruments such as the piano, the guitar, or the bass guitar, the equation describing the vibration of an ideal flexible string is

extended by a restoring force caused by the string stiffness [7]. Due to dispersive wave propagation within the vibrating string, the effect of inharmonicity occurs, i.e., the purely harmonic frequency relationship of an ideal string is distorted and the harmonic frequencies are stretched towards higher values as

$$f_k = k f_0 \sqrt{1 + \beta k^2};\ k \geq 1 \tag{1}$$

with k being the harmonic index of each overtone and f_0 being the fundamental frequency. The inharmonicity coefficient β depends on different properties of the vibrating string such as Young's Modulus E, the radius of gyration K, the string tension T, the cross-sectional area S, as well as the string length L. With the string length being approximately constant for all strings of the bass guitar and the electric guitar, the string diameter usually varies from 0.45 mm to 1.05 mm for electric bass and from 0.1 mm to 0.41 mm for electric guitar[2]. The string tension T is proportional to the square of the fundamental frequency of the vibrating string. Järveläinen et al. performed different listening tests to investigate the audibility of inharmonicity towards humans [12]. They found that the human audibility threshold for inharmonicity increases with increasing fundamental frequency.

Hodgekinson et al. observed a systematic time-dependence of the inharmonicity coefficient if the string is plucked hard [10]. The authors found that β does not remain constant but increases over time for an acoustic guitar note. In contrast, for a piano note, no such behavior was observed. In this paper, we aim to estimate β on a frame-level and do not take the temporal evolution of β into account.

Different methods have been applied in the literature to extract the inharmonicity coefficient such as the cepstral analysis, the harmonic product spectrum [8], or inharmonic comb-filter [9]. For the purpose of sound synthesis, especially for physical modeling of string instruments, inharmonicity is often included into the synthesis models in order to achieve a more natural sound [24]. The inharmonicity coefficient of different instruments was analyzed as a distinctive feature in different Music Information Retrieval tasks such as instrument recognition and music transcription.

3.2 Hybrid Approaches & Visual Approaches

Different methods for estimating the fretboard position from guitar recordings were presented in the literature that include analysis methods from computer vision as a multi-modal extension of audio-based analysis.

A combined audio and video analysis was proposed by Hybryk and Kim to estimate the fretboard position of chords that were played on an acoustic guitar [11]. The goal of this paper was to first identify a played chord on the guitar regarding its "chord style", i.e., their root note and musical mode such as minor or major. For this purpose, the Specmurt [22] algorithm was used for

[2] These values correspond to commonly used string gauges.

spectral analysis in order to estimate a set of fundamental frequency candidates that can be associated to different note pitches. Based on the computed "chord style" (e.g., E minor), the "chord voicing" was estimated by tracking the spatial position of the hand on the instrument neck. The chord voicing is similar to the chord fingering as described in [2].

Another multi-modal approach for transcribing acoustic guitar performances was presented by Paleari et al. in [19]. In addition to audio analysis, the visual modality was analyzed to track the hand of the guitar player during his performance to estimate the fretboard position. The performing musicians were recorded using both two microphones and a digital video camera. The fretboard was first detected and then spatially tracked over time.

Other approaches solely used computer vision techniques for spatial transcription. Burns and Wanderley presented an algorithm for real-time finger-tracking in [4]. They used *attached cameras* on the guitar in order to get video recordings of the playing hand on the instrument neck. Kerdvibulvech and Saito used a stereo-camera setup to record a guitar player in [13]. Their system for finger-tracking requires the musician to wear *colored fingertips*. The main disadvantage of all these approaches is that both the attached cameras as well as the colored fingertips are unnatural for the guitar player. Therefore, they likely limit and impede the musician's expressive gestures and playing style.

Enhanced music instruments are equipped with additional sensors and controllers in order to directly measure the desired parameters instead of estimating them from the audio or video signal. On the one hand, these approaches lead to a high detection accuracy. On the other hand, these instrument extensions are obtrusive to the musicians and can affect their performance on the instrument [11]. In contrast to regular electric guitar pickups, *hexaphonic pickups* separately capture each vibrating string. In this way, spectral overlap between the string signals is avoided, which allows a fast and robust pitch detection with very low latency and very high accuracy, as shown for instance by O'Grady and Rickard in [18].

4 New Approach

Fig. 2 provides an overview over the string classification algorithm proposed in this paper. All processing steps are explained in detail in the next sections.

4.1 Spectral Modeling

Non-parametric spectral estimation methods such as the Periodogram make no explicit assumption on the type of signal that is analyzed. In order to obtain a sufficiently high frequency resolutions for a precise f_0-detection, relatively large time frames of data samples are necessary in order to compensate the spectral leakage effect, which is introduced by windowing the signal into frames. In contrast to the percussive nature of its short attack part (between approx. 20 ms and 40 ms), the decay part of a plucked string note can be modeled by a sum of decaying sinusoidal components. Their frequencies have a nearly perfectly

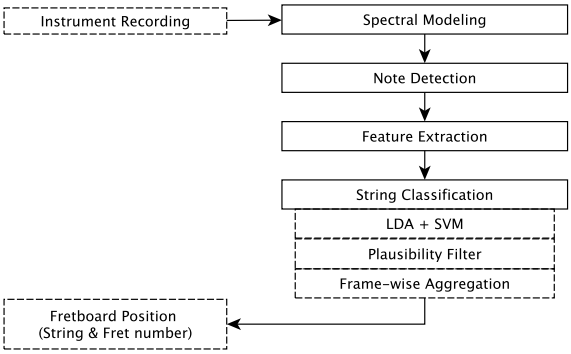


Fig. 2: Algorithm overview

harmonic relationship. Since the strings of the bass guitar and the electric guitar have a certain amount of stiffness, the known phenomenon of inharmonicity appears (compare Sect. 3.1).

Parametric spectral estimation techniques can be applied if the analyzed signal can be assumed to be generated by a known model. In our case, the power spectral density (PSD) $\Phi(\omega)$ can be modeled by an auto-regressive (AR) filter such as

$$\Phi(\omega) \approx \Phi_{AR}(\omega) = \sigma^2 \left| \frac{1}{1 + \sum_{l=1}^{p} a_l e^{-jl\omega}} \right|^2 \tag{2}$$

with σ^2 denoting the process variance, p denoting the model order, and $\{a_l\} \in \mathbb{R}^{p+1}$ being the filter coefficients. Since auto-regressive processes are closely related to linear prediction (LP), both a *forward prediction error* and a *backward prediction error* can be defined to measure the predictive quality of the AR filter. We use the *least-squares method* (also known as *modified covariance method*) for spectral estimation. It is based on a simultaneous least-squares minimization of both prediction errors with respect to all filter coefficients $\{a_l\}$. This method has been shown to outperform related algorithms such as the Yule-Walker method, the Burg algorithm, and the covariance method (See [17] for more details). The size of the time frames N is only restricted by the model order as $p \leq 2N/3$.

First, we down-sample the signals to $f_s = 5.5$ kHz for the bass guitar samples and $f_s = 10.1$ kHz for the electric guitar samples. This way, we can detect the first 15 harmonics of each note within the instrument pitch ranges, which is necessary for the subsequent feature extraction as explained in Sect. 4.2. In Fig. 3, the estimated AR power spectral density for a bass guitar sample (E1) as well as the estimated partials are illustrated. Since we only focus on isolated instrument samples here, we assume the fundamental frequency f_0 to be known in advance. The separate evaluation of fundamental frequency estimation is not within the scope of this paper.

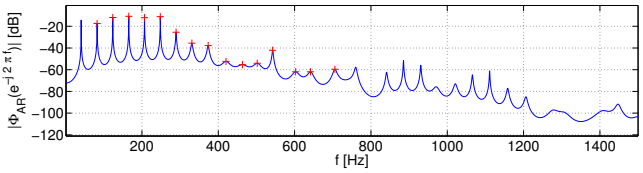


Fig. 3: Estimated AR power spectral density for the bass guitar sample with pitch E1 ($f_0 = 44.1Hz$). The estimated first 15 partials are indicated with red crosses.

By using overlapping time frames with a block-size of $N = 256$ and a hop-size of $H = 64$, we apply the spectral estimation algorithm to compute frame-wise estimates of the filter coefficients $\{a_l(n)\}$ in the frames that are selected for analysis (compare Sect. 4.2). In order to estimate the harmonic frequencies $\{f_k\}$, we first compute the pole frequencies of the AR filter by computing the roots of the numerator in Eqn. (2). Then, we assign one pole frequency to each harmonic according to the highest proximity to its theoretical frequency value as computed using Eqn. (1).

4.2 Feature Extraction

Note Detection In Sect. 4.1, we discussed that notes played on the bass guitar and the guitar follow a signal model of decaying sinosoidal components, i.e., the partials. In this section, we discuss how we extract audio features that capture the amplitude and frequency characteristics. We first detect the first frame shortly after the note attack part of the note is finished and the harmonic decay part begins. As mentioned in Sect. 4.1, signal frames with a percussive characteristic are indicated by high values of the process variance $\sigma^2(t)$ obtained the AR spectral estimation. We found that time frames after

$$t^\star = \arg\max_t \sigma^2(t) \tag{3}$$

are suitable for feature extraction. If the aggregation of multiple frame-wise results is used, we extract features in the first 5 frames after $t^\star$.

Inharmonicity estimation In each analyzed frame, we estimate the discrete frequencies f_k of the first 15 partials. Then, we estimate the inharmonicity coefficient β_k as follows. From Eq. (1), we obtain

$$(f_k/f_0)^2 = k^2 + \beta k^4 \tag{4}$$

We use polynomial curve fitting to approximate the left-hand side of Eq. (4) by a polynomial function of order 4 as

$$(f_k/f_0)^2 \approx \sum_{i=0}^{4} p_i k^i \tag{5}$$

Feature	Feature dimension
Inharmonicity coefficient $\hat{\beta}$	1
Relative partial amplitudes $\{\hat{a}_{r,k}\}$	15
Statistics over $\{\hat{a}_{r,k}\}$	8
Normalized partial frequency deviations $\{\Delta\hat{f}_{norm,k}\}$	15
Statistics over $\{\Delta\hat{f}_{norm,k}\}$	8
Partial amplitude slope $\hat{s}_a$	1
All features	$\sum = 48$

Table 1: Overview of all applied audio features.

and use the coefficient p_4 as an estimate of the inharmonicity coefficient β:

$$\hat{\beta} \approx p_4 \tag{6}$$

Partial-based Features In addition to the inharmonicity coefficient β, we compute various audio features that capture the amplitude and frequency characteristics of the first 15 partials of a note. First, we compute the relative amplitudes

$$\{\hat{a}_{r,k}\} = \{a_k/a_0\} \tag{7}$$

of the first 15 partials related to the amplitude of the fundamental frequency. Then, we approximate the relative partial amplitude values $\{\hat{a}_{r,k}\}$ as a linear function over k as

$$\hat{a}_{r,k} \approx p_1 k + p_0 \tag{8}$$

by using linear regression. We use the feature $\hat{s}_a = p_1$ as estimate of the *spectral slope* towards higher partial frequencies.

Based on the estimated inharmonicity coefficient $\hat{\beta}$ and the fundamental frequency f_0, we compute the theoretical partial frequency values $\{f_{k,theo}\}$ of the first 15 partials based on Eq. (1) as

$$f_{k,theo} = k f_0 \sqrt{1 + \hat{\beta} k^2}. \tag{9}$$

Then, we compute the deviation between the theoretical and estimated partial frequency values and normalize this difference value as

$$\Delta\hat{f}_{norm,k} = \frac{f_{k,theo} - \hat{f}_k}{\hat{f}_k}. \tag{10}$$

Again, we compute $\{\Delta\hat{f}_{norm,k}\}$ for the first 15 partials and use them as features. In addition, we compute the statistical descriptors maximum value, minimum value, mean, median, mode (most frequent sample), variance, skewness, and kurtosis over both $\{\hat{a}_{r,k}\}$ and $\{\Delta\hat{f}_{norm,k}\}$. Tab. 1 provides an overview over all dimensions of the feature vectors.

4.3 Estimation Of The Fretboard Position

String Classification In order to automatically estimate the fretboard position from a note recording, we first aim to estimate the string number n_s. Therefore, we compute the 48-dimensional feature vector $\{x_i\}$ as described in the previous section. We use Linear Discriminant Analysis (LDA) to reduce the dimensionality of the feature space to $N_d = 3$ dimensions for bass guitar and to $N_d = 5$ dimensions for guitar[3]. Then we train a Support Vector Machine (SVM) classifier using a Radial Basis Function (RBF) kernel with the classes defined by notes played on each string. SVM is a binary discriminative classifier that attempts to find an optimal decision plane between feature vectors of the different training classes [25]. The two kernel parameters C and γ are optimized based on a three-fold grid search. We use the LIBSVM library for our experiments [5].

The SVM returns probabilities $\{p_i\}$ to assign unknown samples to each string class. We estimate the string number $\hat{n}_s$ as

$$\hat{n}_s = \arg\max_i \{p_i\}. \tag{11}$$

We derive the the fret number $\hat{n}_f$ from the estimated string number $\hat{n}_s$ by using knowledge on the instrument tuning as follows. The common tuning of the bass is E1, A2, D2, and G2; the tuning of the guitar is E2, A2, D3, G3, B3, and E3. The string tunings can be directly translated into a vector of corresponding MIDI pitch values as $\{P_T\} = [28, 33, 38, 43]$ and $\{P_T\} = [40, 45, 50, 55, 59, 64]$, respectively.

In order to derive the fret number $\hat{n}_s$, we first obtain the MIDI pitch value P that corresponds to the fundamental frequency f_0 as

$$P = \lfloor 12 \log_2(f_0/440) - 69 \rfloor \tag{12}$$

Given the estimated string number $\hat{n}_s$, the fret number can be computed as

$$\hat{n}_f = P - P_T(\hat{n}_s). \tag{13}$$

A fret number of $\hat{n}_f = 0$ indicates that a note was played by plucking an open string.

Plausibility Filter As mentioned earlier, most note pitches within the frequency range of both the bass guitar and the guitar can be played on either one, two, or three different fret positions on the instrument neck. The pitch ranges are E2 to G3 for the bass guitar and E3 to E5 for the electric guitar considering a fret range up to the 12th fret position. Based on knowledge about the instrument tunings, we can derive a set of MIDI pitch values that can be played on each string. Therefore, for each estimated MIDI pitch value $\hat{P}$, we can derive a list of strings, where this note can theoretically be played on. If the plausibility filter is used, we set the probability values in $\{p_i\}$ of all strings, where this note can not be played on to 0 before estimating the string number as shown in Eq. (11).

[3] The number of dimensions N_d is chosen as $N_d = N_{strings} - 1 \equiv N_{classes} - 1$.

Aggregation of multiple classification results If the result aggregation is used, we sum up all class probability values $\{p_i\}$ over 5 adjacent frames. Then we estimate the string number as shown in Eq. (11) over the accumulated probability values.

5 Evaluation & Results

5.1 Dataset

For the evaluation experiments, we use a dataset of 1034 audio samples. These samples are isolated note recordings, which were taken from the dataset previously published in [23].[4] The samples were recorded using two different bass guitars and two different electric guitars, each played with two different plucking styles (plucked with a plectrum and plucked with the fingers) and recorded with two different pick-up settings (either neck pick-up or body pick-up).

5.2 Experiments & Results

Experiment 1: Feature Selection for String Classification In this experiment, we aim to identify the most discriminant features for the automatic string classification task as discussed in Sect. 4.3. Therefore, we apply the feature selection algorithm Inertia Ratio Maximization using Feature Space Projection (IRMFSP) [15, 20] to all feature vectors and the corresponding class labels separately for both instrument. In Tab. 2, the five features that are first selected by the IRMFSP algorithm are listed for the bass guitar and the electric guitar.

The features $\Delta\hat{f}_{norm}$, $\hat{\beta}$, and $\hat{a}_{r,k}$ as well as the derived statistic measures were selected consistently for both instruments. These features measure frequency and amplitude characteristics of the partials and show high discriminative power between notes played on different strings independently of the applied instrument. The boxplots of the two most discriminative features $\Delta f_{norm,9}$ for bass and $\Delta f_{norm,15}$ for guitar are illustrated separately for each instrument string in Fig. 4.

Since the deviation of the estimated harmonic frequencies from their theoretical values apparently carries distinctive information to discern between notes on different instrument strings, future work should investigate, if Eq. (1) could be extended by higher order polynomial terms in order to better fit to the estimated harmonic frequency values.

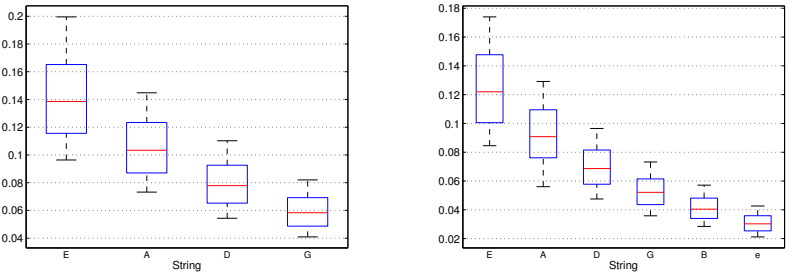
Experiment 2: String Classification in different conditions In this experiment, we aim to investigate how the performance of the automatic string classification algorithm is affected by

[4] This dataset contains isolated notes from bass guitar and electric guitar processed with various audio effects. In this work, only the non-processed note recordings were used.

Rank	Bass Guitar	Electric Guitar
1	$\Delta\hat{f}_{norm,9}$	$\Delta\hat{f}_{norm,15}$
2	$\hat{\beta}$	mean$\{\hat{a}_{r,k}\}$
3	$\Delta\hat{f}_{norm,3}$	var$\{\Delta\hat{f}_{norm,k}\}$
4	var$\{\Delta\hat{f}_{norm,k}\}$	max$\{\hat{a}_{r,k}\}$
5	$\hat{a}_{r,4}$	skew$\{\Delta\hat{f}_{norm,k}\}$

Table 2: Most discriminative audio features for the string classification task as discussed in Sect. 5.2. Features are given in order as selected by the IRMFSP algorithm.

- the separation of the training and test set according to the applied instrument, playing technique, and pick-up setting,
- the instrument / the number of string classes,
- the use of a plausibility filter (compare Sect. 4.3),
- and the use of a aggregation of multiple classification results for each sample (compare Sect. 4.3).



(a) Boxplot of feature $\Delta f_{norm,9}$ for bass.

(b) Boxplot of feature $\Delta f_{norm,15}$ for guitar.

Fig. 4: Boxplots of the two most discriminative features for bass guitar and electric guitar.

The different conditions are illustrated in Tab. 3 for the bass guitar and in Tab. 4 for the electric guitar. The colums “Separated instruments”, “Separated playing techniques”, and “Separated pick-up setting” indicate which criteria were applied to separate the samples from training and test set in each configuration. The fifth and sixth column indicate whether the plausibility filter and the frame result aggregation were applied. In the seventh column, the number of folds for the configuration 1.6 and 2.6 and the number of permutations for the remaining configurations are given. The evaluation measures precision, recall, and F-measure were always averaged over all permutations and all folds, respectively.

Experiment	Separated instruments	Separated playing techniques	Separated pick-up settings	Plausibility filter (see Sect. 4.3)	Result aggregation over 5 frames (see Sect. 4.3)	No. of Permutations$^\diamond$ / No. of CV folds*	Precision $\bar{P}$	Recall $\bar{R}$	F-Measure $\bar{F}$
1.1.a	x					$2^\diamond$	.85	.85	.85
1.1.b	x			x		$2^\diamond$	.87	.87	.87
1.1.c	x			x	x	$2^\diamond$	.78	.78	.78
1.2.a	x	x				$8^\diamond$	.86	.86	.86
1.2.b	x	x		x		$8^\diamond$	.87	.87	.87
1.2.c	x	x		x	x	$8^\diamond$	.88	.88	.88
1.3.a		x	x			$8^\diamond$	.57	.50	.49
1.3.b		x	x	x		$8^\diamond$	.71	.69	.69
1.3.c		x	x	x	x	$8^\diamond$	.88	.88	.88
1.4.a		x				$8^\diamond$	.60	.54	.54
1.4.b		x		x		$8^\diamond$	.73	.71	.72
1.4.c		x		x	x	$8^\diamond$	**.93**	**.93**	**.93**
1.5.a			x			$8^\diamond$	.62	.55	.54
1.5.b			x	x		$8^\diamond$	.74	.71	.71
1.5.c			x	x	x	$8^\diamond$	.92	.92	.92
1.6.a						$10^\star$	.92	.92	.92
1.6.b				x		$10^\star$	.93	.93	.93
1.6.c				x	x	$10^\star$	**.93**	**.93**	**.93**

Table 3: Mean Precision $\bar{P}$, mean Recall $\bar{R}$, and mean F-Measure $\bar{F}$ for different evaluation conditions (compare Sect. 5.2) for the bass guitar.

After the training set and the test set are separated, the columns of the training feature matrix were first normalized to zero mean and unit variance. The mean vector and the variance vector were kept for later normalization of the test data. Subsequently, the normalized training feature matrix is used to derive the transformation matrix via LDA. We chose $N = N_{Strings} - 1$ as number of feature dimensions. The SVM model is then trained using the projected training feature matrix and a two-dimensional grid search is performed to determine the optimal parameters C and γ as explained in Sect. 4.3. For the configurations 1.6 and 2.6, none of the criteria to separate the training and the test set was applied. Instead, here we used a 10-fold cross-validation and averaged the precision, recall, and F-measure over all folds.

The results shown in Tab. 3 and Tab. 4 clearly show that both the plausibility filter as well as the result aggregation step significantly improve the classification results in most of the investigated configurations. Furthermore, we can see that the separation of training and test samples according to instrument, playing technique, and pick-up setting has a strong influence on the achievable classification performance. In general, the results obtained for the bass guitar and the electric guitar show the same trends. We obtain the highest classification

Experiment	Separated instruments	Separated playing techniques	Separated pick-up settings	Plausibility filter (see Sect. 4.3)	Result aggregation over 5 frames (see Sect. 4.3)	No. of Permutations$^{\diamond}$ / No. of CV folds*	Precision $\bar{P}$	Recall $\bar{R}$	F-Measure $\bar{F}$
2.1.a	x					$2^{\diamond}$	.64	.64	.63
2.1.b	x			x		$2^{\diamond}$	.70	.70	.70
2.1.c	x			x	x	$2^{\diamond}$	.76	.75	.75
2.2.a	x	x				$8^{\diamond}$	.69	.69	.68
2.2.b	x	x		x		$8^{\diamond}$	.71	.71	.70
2.2.c	x	x		x	x	$8^{\diamond}$	.78	.77	.77
2.3.a		x	x			$8^{\diamond}$	.61	.57	.56
2.3.b		x	x	x		$8^{\diamond}$	.68	.66	.66
2.3.c		x	x	x	x	$8^{\diamond}$	.74	.74	.73
2.4.a		x				$8^{\diamond}$	.64	.61	.60
2.4.b		x		x		$8^{\diamond}$	.71	.69	.69
2.4.c		x		x	x	$8^{\diamond}$	.80	.79	.79
2.5.a			x			$8^{\diamond}$	.69	.65	.65
2.5.b			x	x		$8^{\diamond}$	.74	.72	.72
2.5.c			x	x	x	$8^{\diamond}$	.84	.84	.84
2.6.a						$10^{\star}$	.72	.69	.70
2.6.b				x		$10^{\star}$	.81	.81	.81
2.6.c				x	x	$10^{\star}$	**.90**	**.90**	**.90**

Table 4: Mean Precision $\bar{P}$, mean Recall $\bar{R}$, and mean F-Measure $\bar{F}$ for different evaluation conditions (compare Sect. 5.2) for the electric guitar.

scores—$\bar{F} = .93$ for the bass guitar (4 classes) and $\bar{F} = .90$ for the electric guitar (6 classes)—for the configurations 1.6 and 2.6. These results indicate that the presented method can be successfully applied in different application tasks that require an automatic estimation of the played instrument string. In contrast to [16], we did not make use any knowledge about the musical context such as derived from a musical score.

We performed a baseline experiment separately for both instruments using Mel Frequency Cepstral Coefficients (MFCC) as features as well as LDA and SVM for feature space transformation and classification, respectively (compare Sect. 4.3). The same experimental conditions as in configuration 1.6. and 2.6. (see Sect. 5.2) were used. The classification results were performed and evaluated on a frame level. A 10-fold stratified cross-validation was applied and the results were averaged over all folds. We achieved classification scores of $\bar{F} = .46$ for the bass guitar and $\bar{F} = .37$ for electric guitar.

6 Conclusions

In this paper, we performed several experiments towards the automatic classification of the string number from given isolated note recordings. We presented a selection of audio features that can be extracted on a frame-level. In order to improve the classification results, we first apply a plausibility filter to avoid non-meaningful classification results. Then, we use an aggregation of multiple classification results that are obtained from adjacent frames of the same note. Highest string classification scores of $\bar{F} = .93$ for the bass guitar (4 string classes) and $\bar{F} = .90$ for the electric guitar (6 string classes) were achieved. As shown in a baseline experiment, classification systems based on commonly-used audio features such as MFCC were clearly outperformed for the given task.

7 Acknowledgements

The author likes to thank Michael Stein for the use of his data set. The Thuringian Ministry of Economy, Employment and Technology supported this research by granting funds of the European Fund for Regional Development to the project *Songs2See*[5], enabling transnational cooperation between Thuringian companies and their partners from other European regions.

References

1. J. Abeßer, C. Dittmar, and G. Schuller. Automatic Recognition and Parametrization of Frequency Modulation Techniques in Bass Guitar Recordings. In *Proc. of the 42nd Audio Engineering Sociery (AES) International Conference on Semantic Audio*, pages 1–8, Ilmenau, Germany, 2011.
2. A. M. Barbancho, A. Klapuri, L. J. Tardón, and I. Barbancho. Automatic Transcription of Guitar Chords and Fingering from Audio. *IEEE Transactions on Audio, Speech, and Language Processing*, pages 1–19, 2011.
3. I. Barbancho, A. M. Barbancho, L. J. Tardón, and S. Sammartino. Pitch and Played String Estimation in Classic and Acoustic Guitars. In *Proceedings of the 126th Audio Engineering Society (AES) Convention, Munich, Germany*, 2009.
4. A. Burns and M. Wanderley. Visual Methods for the Retrieval of Guitarist Fingering. In *Proceedings of the 2006 International Conference on New Interfaces for Musical Expression (NIME06)*, pages 196–199, Paris, France, 2006.
5. C.-C. Chang and C.-J. Lin. LIBSVM : A Library for Support Vector Machines. Technical report, Department of Computer Science, National Taiwan University, Taipei, Taiwan, 2011.
6. M. G. Christensen and A. Jakobsson. *Multi-Pitch Estimation*. Synthesis Lectures on Speech and Audio Processing. Morgan & Claypool Publishers, 2009.
7. N. H. Fletcher and T. D. Rossing. *The Physics Of Musical Instruments*. Springer, New York, London, 2 edition, 1998.
8. A. Galembo and A. Askenfelt. Measuring inharmonicity through pitch extraction. *Speech Transmission Laboratory. Quarterly Progress and Status Reports (STL-QPSR)*, 35(1):135–144, 1994.

[5] `http://www.songs2see.net`

9. A. Galembo and A. Askenfelt. Signal representation and estimation of spectral parameters by inharmonic comb filters with application to the piano. *IEEE Transactions on Speech and Audio Processing*, 7(2):197–203, 1999.
10. M. Hodgkinson, J. Timoney, and V. Lazzarini. A Model of Partial Tracks for Tension-Modulated Steel-String Guitar Tones. In *Proc. of the 13th Int. Conference on Digital Audio Effects (DAFX-10), Graz, Austria*, number 1, pages 1–8, 2010.
11. A. Hrybyk and Y. Kim. Combined Audio and Video for Guitar Chord Identification. In *Proc. of the 11th International Society for Music Information Retrieval Conference (ISMIR), Utrecht, Netherlands*, number Ismir, pages 159–164, 2010.
12. H. Järveläinen, V. Välimäki, and M. Karjalainen. Audibility of the timbral effects of inharmonicity in stringed instrument tones. *Acoustics Research Letters Online*, 2(3):79, 2001.
13. C. Kerdvibulvech and H. Saito. Vision-Based Guitarist Fingering Tracking Using a Bayesian Classifier and Particle Filters. *Advances in Image and Video Technology*, pages 625–638, 2007.
14. A. Klapuri. Multipitch analysis of polyphonic music and speech signals using an auditory model. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):255–266, 2008.
15. H. Lukashevich. Feature selection vs. feature space transformation in automatic music genre classification tasks. In *Proc. of the AES Convention*, 2009.
16. A. Maezawa, K. Itoyama, T. Takahashi, T. Ogata, and H. G. Okuno. Bowed String Sequence Estimation of a Violin Based on Adaptive Audio Signal Classification and Context-Dependent Error Correction. In *Proc. of the 11th IEEE International Symposium on Multimedia (ISM2009)*, pages 9–16, 2009.
17. S. L. Marple. *Digital Spectral Analysis With Applications*. Prentice Hall, Australia, Sydney, 1987.
18. P. D. O'Grady and S. T. Rickard. Automatic Hexaphonic Guitar Transcription Using Non-Negative Constraints. In *Proc. of the IET Irish Signals and Systems Conference (ISSC)*, pages 1–6, Dublin, Ireland, 2009.
19. M. Paleari, B. Huet, A. Schutz, and D. Slock. A Multimodal Approach to Music Transcription. In *Proc. of the 15th IEEE International Conference on Image Processing (ICIP)*, pages 93–96, 2008.
20. G. Peeters and X. Rodet. Hierarchical gaussian tree with inertia ratio maximization for the classification of large musical instruments databases. In *Proc. of the Int. Conf. on Digital Audio Effects (DAFx)*, London, UK, 2003.
21. H. Penttinen and J. Siiskonen. Acoustic Guitar Plucking Point Estimation in Real Time. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 209–212, 2005.
22. S. Saito, H. Kameoka, K. Takahashi, T. Nishimoto, and S. Sagayama. Specmurt Analysis of Polyphonic Music Signals. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(3):639–650, Feb. 2008.
23. M. Stein, J. Abeßer, C. Dittmar, and G. Schuller. Automatic Detection of Audio Effects in Guitar and Bass Recordings. In *Proceedings of the 128th Audio Engineering Society (AES) Convention, London, UK*, 2000.
24. V. Välimäki, J. Pakarinen, C. Erkut, and M. Karjalainen. Discrete-time modelling of musical instruments. *Reports on Progress in Physics*, 69(1):1–78, Jan. 2006.
25. V. N. Vapnik. *Statistical learning theory*. Wiley New York, 1998.

Improving Beat Tracking in the presence of highly predominant vocals using source separation techniques: Preliminary study

José R. Zapata and Emilia Gómez

Music Technology Group
Universitat Pompeu Fabra
{joser.zapata,emilia.gomez}@upf.edu

Abstract. The automatic beat tracking from audio is still an open research task in the Music Information Retrieval (MIR) community. The goal of this paper is to show and discuss a work-in-progress of how audio source separation can be used for improving beat tracking estimations in difficult cases of music audio signal with highly predominant vocals. The audio source separation using FASST (Flexible Audio Source Separation Toolbox) had an average improvement of beat tracking of {14,15%, 17,74%} in the F-measure and {14,21%, 25,70%} in the Amlt of Klapuri and Degara systems respectably in a dataset of 20 songs excerpt.

Keywords: Beat tracking, Source separation, Predominant voice

1 Introduction

The task of Beat tracking is related to the detection of the main pulse beat, defined as "one of a series of regularly recurring, precisely equivalent stimuli" [1]. For Western music, a hierarchical metrical structure is found in different time scales, and the most common ones are: the tatum period, defined as "a regular time division that mostly coincides with all note onsets"; and the tactus period (the perceptually most prominent period), defined as the rate at which most people would regularly tap their feet, hands or finger in time following the music.

Beat is a relevant audio descriptor of a piece of music, which represents the speed of the piece under study. For that reason, much research within the Music Information Retrieval (MIR) community has been devoted to finding ways to automate its extraction and many algorithms have been proposed. Beat tracking algorithms have been used in different application contexts, such as music retrieval, cover detection, playlist generation, and beat synchronization for audio mixing, structural analysis and score alignment. Many approaches for beat tracking have been proposed, and some efforts have been devoted to their quantitative comparisons to find other ways to emphasize and detect the rhythm accents in music, but it's not still clear in which kind of music or interpretations the beat trackers have problems to detect the beats.

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

A recent study in beat tracking difficulty [2] presented a technique for estimating the degree of difficulty of musical excerpts in beat tracking based on the mutual agreement between a committee of beat tracking algorithms. In this study an audio dataset was built containing 678 excerpts of 40s length from various musical styles such as classical, chanson, jazz, folk and flamenco. In this study difficult cases for beat tracking songs with strong and expressive voice were found. Even with a stable accompaniment, beat trackers encountered problems.

The goal of this paper is to present and discuss a work-in-progress of the improvement of beat tracking estimation in difficult cases with highly predominant vocals, using FASST (Flexible Audio Source Separation Toolbox). Based on the evidence, a discussion of the results and ideas for future work are presented.

This paper is structured as follows. First, we present current challenges for beat tracking, followed by the hypothesis of the experiment. Second, Each part of the evaluated system is briefly explained. Third, we present the results of each beat tracking experiment. Finally, we provide some discussions, limitations, future work and conclusions of this study.

2 Experiment Hypothesis

The hypothesis of this experiment originated from previous research on: automatic beat tracking with percussive/ harmonic separation [3] and tempo estimation that uses source separation [4] or percussive/harmonic separation[5] to improve tempo detection. Based on this research, a source separation technique is proposed to improve beat tracking in difficult cases with highly predominant vocals and quiet accompaniment.

3 Experimental Framework

The main goal of the experiment is to evaluate if audio source separation techniques improve the beat tracking systems. The experiment consists of an evaluation of two beat tracking algorithms on 20 audio song excerpts (highly predominant vocals) before and after a process of source separation.

3.1 Audio Beat Trackers

Two different systems were used for this experiment:

1. The Matlab implementation of the well-known Audio Beat tracking system by Anssi Klapuri [6], which uses the differentials of loudness in 36 frequency subbands as audio features which are then combined in four signals. These signals measure the degree of musical accentuacion over time. The pulse induction block is a bank comb filter. The algorithm estimates the tatum, the beat and the measure through probabilistic modeling the relationships and temporal evolutions.

2. The Matlab implementation of Degara's beat tracker by Norberto Degara [7], analyzes the input musical signal based on complex spectral difference

method and extracts a beat phase and a beat period salience observation signal, with this info estimates the time between consecutive beat events and exploits both beat and non-beat information by explicitly modeling non-beat states. In addition to the beat times, a measure of the expected accuracy of the estimated beats is provided. The quality of the observations used for beat tracking are measured and the reliability of the beats is automatically calculated. The accuracy of the beat estimations are predicted by a k-nearest neighbor regression algorithm.

3.2 Audio Source Separation

The Matlab software tool named Flexible Audio Source Separation Toolbox (FASST) [10] we used as a source separation tool for the experiment. The framework can incorporate prior information about the audio signal. The basic example (EXAMPLE_prof_rec_sep_drums_bass_melody.m) contains information allowing the separation of the following four sources: Bass, Drums, melody (singing voice or leading melodic instrument) and remaining sounds (other).
The Framework FASST is available in `http://bass-db.gforge.inria.fr/fasst/`

3.3 Music Material

The audio files used in the experiment are a subset of 20 excerpts from the databases used in [2]. It consists of difficult song cases of audio beat tracking with highly predominant vocals and the format is the same for all: mono, linear PCM, 44100 Hz sampling frequency, 16 bits resolution. Each excerpt has ground truth annotations of the beats as described in [2]. The artist and the name of each song are in Table 1 and Table 2.

3.4 Evaluation methods

We contrasted the beat trackers output from the original excerpts ans the output of the source separation method. The evaluation techniques considered in this study are:

F-measure [8] : Beats are considered accurate if they fall within a 70ms tolerance window around annotations. Accuracy in a range from 0% to 100% is measured as a function of the number of true positives, false positives and false negatives.

AMLt [9]: A continuity-based method, where beats are accurate when consecutive beats fall within tempo-dependent tolerance windows around successive annotations. Beat sequences are also accurate if the beats occur on the off-beat, or are tapped at double or half the annotated tempo. The range of values for AMLt is 0% to 100%.

It's important to note that F-measure can increase either due to and increase of tru positives or decrease of false positives or negatives. The Amlt measure improvement can be due to the estimation of true positives in different metrical levels, and continuity is not required.

4 Results

Table 1 and Table 2 present the evaluation results of F-measure and Amlt evaluation for Klapuri and Degara beat tracking algorithms respectively from the original excerpts and the source separation output files.

The average result for the original excerpts of Klapuri algorithm is {39,61%, 39,02%} for F-measure and Amlt respectively. Taking only the best beat tracking result from the separated signals per each song, the average resultincreases to {50,43%, 51,97%} for F-measure and Amlt respectively.

For Degara method, the average result for the original excerpts is equal to {33,6%, 28,6%} for F-measure and Amlt respectively. Considering only the best beat tracking result from the separated signals per each song, the average result increases to {45,71%, 47,78%} for F-measure and Amlt respectively.

Results of Klapuri beat tracker using source separation improved 95% on the dataset at least in one measure. F-measure values in 80% of the dataset in a range of {0,3%, 39,67%} (50% on the Bass) and Amlt values in 90% of the dataset in a range of {1,49%, 37,01%} (33,33% on the Bass). Results of Degara beat tracker using source separation improved 85% on the dataset at least in one measure. F-measure values in 75% of the dataset in a range of {1,6%, 46%} (53,33% on the Bass) and Amlt values in 80% of the dataset in a range of {0,3%, 72,95%} (50% on the Bass).

5 Discussion, Limitations and Future work

In the presented experiment we show that, most of the time, beat tracking estimations can be improved by means of source separation techniques in highly predominant vocal songs, although the expressiveness of the voice such as vibrato, rubato, etc, can difficult beat tracking. In future work we will also consider a low latency voice elimination technique (de-soloing) [11] as an alternative option.

5.1 Source Separation

The FASST source separation tools allow source separation without collecting prior information about the input audio signal. One problem is the computational time because it takes more than 20 minutes to process each audio signal. One limitation for source separation is the few implemented and tested systems to use for academic research and implementing low latency algorithms is still a research challenge. For future experiments different source separation systems had to be evaluated to determine the best alternative for our problem.

From the evaluation results Bass output had better results but is not clear which of the four outputs from the source separation is better to use in all the cases, as it depends on the instruments present in the song. A rhythm strength level measure per signal could be used for this purpose, so that we would apply the beat tracking algorithm in the output signal with higher rhythm strength. One open issue is how to combine the beat tracking estimations from the different sources of the same song to improve beat tracking results.

Artist - Song title	Measure	Original	Melody	Bass	Drums	Other
Joss Stone	**F-measure**	26,51	31,71	**34,04**	29,27	32,10
Dirty Man	**Amlt**	3,08	2,04	**13,85**	2,04	4,17
Edith Piaf	**F-measure**	47,80	42,70	50,91	**53,41**	44,32
La Foule	**Amlt**	22,41	35,48	44,83	**56,67**	56,67
Joss Stone	**F-measure**	22,86	19,13	14,58	**23,16**	23,16
The Chokin' Kind	**Amlt**	9,88	**20,99**	9,09	12,96	11,11
Diana Krall	**F-measure**	18,18	9,26	**32,65**	16,82	8,00
Just The Way You Are	**Amlt**	8,00	8,00	17,33	**22,67**	4,00
Tomwaits	**F-measure**	17,48	40,38	29,03	34,86	**57,14**
The Piano Has Been Drinking	**Amlt**	38,46	41,51	12,68	33,93	**75,47**
Tomwaits	**F-measure**	31,07	30,91	32,65	20,00	**38,46**
Foreign Affair.wav	**Amlt**	18,99	25,32	20,69	8,33	18,99
Joss Stone	**F-measure**	8,33	8,33	15,22	**22,50**	8,33
Understand	**Amlt**	67,35	63,27	0,00	24,56	**75,51**
Tomwaits	**F-measure**	44,44	24,24	**54,35**	14,58	20,45
The One That Got Away	**Amlt**	65,00	26,09	**90,32**	21,21	42,37
Edith Piaf	**F-measure**	28,32	**40,35**	18,18	20,34	21,43
L'Accordeoniste	**Amlt**	13,56	**23,33**	13,43	17,19	8,62
Edith Piaf	**F-measure**	50,00	26,80	**79,12**	28,83	21,05
Correqu' Et Reguyer	**Amlt**	56,63	21,82	**67,35**	31,33	26,42
Edith Piaf	**F-measure**	27,87	19,67	**42,59**	32,73	31,67
Prisonnier De La Tour	**Amlt**	11,34	4,11	**35,59**	16,39	12,37
Edith Piaf	**F-measure**	14,81	22,43	24,30	29,36	**33,64**
Il Pleut	**Amlt**	7,69	14,06	4,71	9,41	**18,75**
Diana Krall	**F-measure**	36,17	15,53	31,11	34,34	31,11
Abandoned Masquerade	**Amlt**	40,00	17,57	**45,90**	30,00	36,07
ABBA	**F-measure**	80,65	77,42	47,62	**93,55**	75,41
The Winner Takes It All	**Amlt**	83,87	87,10	43,75	**96,77**	80,65
Tony Bennett	**F-measure**	21,74	18,60	**42,55**	31,11	24,39
i used to be colourblind	**Amlt**	35,48	6,90	**56,25**	33,33	27,59
Ivor Novello	**F-measure**	17,54	29,51	**32,65**	3,70	18,87
I Can Give You	**Amlt**	14,29	**21,88**	20,00	17,86	13,79
Joe Cocker	**F-measure**	80,28	77,14	28,57	52,35	68,57
That's the way her love is	**Amlt**	85,92	**90,14**	14,44	44,87	94,37
Roberto Goyeneche	**F-measure**	74,29	38,46	67,29	51,92	**78,10**
Ventanita florida	**Amlt**	81,13	40,38	67,27	48,08	81,13
Bruce Springsteen	**F-measure**	87,34	11,45	28,00	82,82	86,34
Thunder Road	**Amlt**	85,34	73,68	9,20	79,82	**86,84**
Meat Loaf	**F-measure**	56,60	39,75	41,10	36,76	52,56
Bat out of hell	**Amlt**	31,97	30,61	25,00	30,65	26,53

Table 1. F-measure and Amlt results for Klapuri beat tracking algorithm

Artist - Song title	Measure	Original	Melody	Bass	Drums	Other
Joss Stone	**F-measure**	36,70	23,93	**46,15**	32,97	26,83
Dirty Man	**Amlt**	38,16	14,29	0,00	**38,46**	3,08
Edith Piaf	**F-measure**	44,32	40,82	40,41	40,21	29,32
La Foule	**Amlt**	30,43	3,75	1,30	30,14	6,67
Joss Stone	**F-measure**	13,46	17,58	**41,07**	32,20	28,57
The Chokin' Kind	**Amlt**	14,29	20,00	**46,91**	35,80	32,94
Diana Krall	**F-measure**	17,02	14,29	**39,25**	20,00	22,86
Just The Way You Are	**Amlt**	7,14	16,67	**46,67**	17,33	21,33
Tomwaits	**F-measure**	34,11	21,24	22,61	33,33	**35,71**
The Piano Has Been Drinking	**Amlt**	10,48	23,33	24,19	26,23	**40,68**
Tomwaits	**F-measure**	36,04	29,63	23,85	21,36	24,00
Foreign Affair.wav	**Amlt**	36,71	32,91	18,99	5,06	17,72
Joss Stone	**F-measure**	17,78	7,84	14,74	5,48	**25,32**
Understand	**Amlt**	17,91	0,00	0,00	28,00	**28,57**
Tomwaits	**F-measure**	27,72	24,49	**52,75**	9,88	25,26
The One That Got Away	**Amlt**	28,17	30,88	**83,61**	6,78	44,62
Edith Piaf	**F-measure**	29,06	21,05	11,97	21,85	14,68
L'Accordeoniste	**Amlt**	15,87	16,67	14,29	**20,00**	16,36
Edith Piaf	**F-measure**	32,08	**38,33**	36,36	20,00	18,00
Correqu' Et Reguyer	**Amlt**	13,25	38,55	**49,40**	14,46	8,62
Edith Piaf	**F-measure**	34,38	32,06	**54,17**	43,56	35,29
Prisonnier De La Tour	**Amlt**	23,71	25,77	**73,47**	46,15	30,19
Edith Piaf	**F-measure**	19,64	18,69	23,21	27,35	**28,57**
Il Pleut	**Amlt**	7,06	4,71	10,59	**21,18**	16,36
Diana Krall	**F-measure**	28,30	17,65	21,95	24,14	24,49
Abandoned Masquerade	**Amlt**	15,58	5,48	**24,56**	0,00	20,29
ABBA	**F-measure**	31,43	32,88	16,67	**77,42**	27,45
The Winner Takes It All	**Amlt**	7,69	0,00	29,41	**80,65**	0,00
Tony Bennett	**F-measure**	20,00	32,65	**38,10**	16,00	17,02
i used to be colourblind	**Amlt**	31,43	44,12	**44,83**	34,29	28,13
Ivor Novello	**F-measure**	57,14	35,29	25,00	**64,52**	34,62
I Can Give You	**Amlt**	44,12	3,45	25,00	**54,55**	4,35
Joe Cocker	**F-measure**	59,15	46,81	**69,01**	32,43	41,42
That's the way her love is	**Amlt**	84,51	71,83	81,69	27,66	36,73
Roberto Goyeneche	**F-measure**	16,36	12,84	37,84	31,48	**59,62**
Ventanita florida	**Amlt**	44,83	52,63	32,20	33,93	**67,31**
Bruce Springsteen	**F-measure**	76,39	39,60	34,04	29,95	55,70
Thunder Road	**Amlt**	70,83	14,16	37,70	13,51	50,00
Meat Loaf	**F-measure**	40,94	43,02	**71,74**	52,24	42,86
Bat out of hell	**Amlt**	29,93	30,61	**40,14**	31,67	31,29

Table 2. F-measure and Amlt results from Degara beat tracking algorithm

5.2 Data

It's important to note that this evaluation has been specifically carried out for difficult beat tracking cases with highly predominant vocals in the audio signal and one limitation is found with these kinds of cases from the beat tracking databases that exist right now with ground truth. For future evaluation, more data with these issues could be collected using an automatic identification system of difficult examples for beat tracking[2] and manually classifying highly predominant vocals cases, or by using an automatic highly predominant vocals detection system.

Most of the source separation algorithms use the spatial information to improve the separation. In this evaluation the datasets are mono audio signals. For future evaluations, it would be good to add some stereo song excepts.

5.3 Beat Tracking

The song excerpt with best improvement of F-measure (13,46% to 41,07%) with Degara algorithm is the sameas the Klapuri has the lowest improvement (22,86% to 23,16%), but the Klapuri algorithm reach better F-measure result for this song excerpt. One limitation of the beat tracking evaluation is the use of different measures to determinate the good performance of the systems. There is no consensus on how to measure with a single value, or which evaluation measure is more reliable for beat tracking proposes.

The Beat tracking in the source separated signals fail when the accompaniment had pauses, tempo changes and the principal metrical level is a musical combination between of all the instruments and the voice (e.g Diana Krall - Abandoned Masquerade).

Another limitation is the lack of methodology to combine the beat tracking results from different algorithms. For future work this evaluation can be performed with more beat trackers to extend the results of the experiment and establish more accurate statements of the advantage of use source separation for improve beat tracking. The evaluation and research of this method can be applied like a pre-process stage in beat tracking.

6 Conclusions

The audio source separation made by FASST algorithm had an average improvement of beat tracking of {14,15%, 17,74%} in the F-measure and {14,21%, 25,70%} in Amlt of Klapuri and Degara systems.

Comparing only the best result from each separated signals per song with the original beat tracking result, the Klapuri and Degara algorithms enhanced the average results in {10,81%, 12,1%} for F-measure and {12,96, 19,18%} for Amlt value respectively.

The Bass output from the source separation enhanced the beat tracking results in the dataset more than the other outputs at least in 50% on F-measure

and 33% on the Amlt for Klapuri and Degara Beat trackers. This is the clearest and common instrument output in most of the songs on the dataset.

Audio source separation could then be used as a pre-process stage for improving beat tracking estimation in difficult songs with highly predominant vocals, without changing the beat tracking algorithm.

Acknowledgments. Thanks to Anssi Klapuri, Norberto Degara and A. Ozerov, E. Vincent and F. Bimbot, the authors of the algorithms of beat tracking and source separation respectively for making their algorithms available for research topics. Matthew Davies, Andre Holzapfel and Fabien Gouyon for the intership support in INESC in Porto. Thanks to Colciencias and Universidad Pontificia Bolivariana (Colombia), Music Technology Group at Universitat Pompeu Fabra, Classical Planet and DRIMS project for the financial support. Robin Motheral for the paper review and Justin Salamon for your helpful recommendations.

References

1. Cooper, G., Meyer, L. B.: The rhythmic structure of music. University Of Chicago Press,Chicago (1960)
2. Holzapfel, A., Davies, M.E.P., Zapata, J.,Oliveira, J.L., Gouyon, F.: On the automatic identification of difficult examples for beat tracking: towards building new evaluation datasets. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP. IEEE Press, Kyoto, Japan (2012)
3. Gkiokas, A., Katsouros, V., Carayannis, G.: ILSP Audio Beat Tracking Algorithm for MIREX 2011. 6th Music Information Retrieval Evaluation eXchange (MIREX). Miami (2011)
4. Chordia, P., Rae, A.: Using source separation to improve tempo detection. In: Proceedings of 10th International Conference on Music Information Retrieval ISMIR, pp. 183–188 (2009)
5. Gkiokas, A., Katsouros, V., Carayannis, G.: ILSP Audio Tempo Estimation Algorithm for MIREX 2011. 6th Music Information Retrieval Evaluation eXchange (MIREX). Miami (2011)
6. Klapuri, A. P., Eronen, A. J., Astola, J. T.: Analysis of the meter of acoustic musical signals. In: IEEE Trans. on Audio, Speech, and Language Processing, vol. 14, no. 1, pp. 342–355 (2006)
7. Degara N., Argones, E., Pena, A., Torres-guijarro, S.,Davies, M.E.P., Plumbley, Mark, D.: Reliability-Informed Beat Tracking of Musical Signals. IEEE Transactions on Audio, Speech and Language Processing, Vol. 20, pp. 290–301 (2012)
8. Dixon,S.: Evaluation of the audio beat tracking system BeatRoot. Journal of New Music Research, vol. 36, pp.39–50 (2007)
9. Hainsworth,S.W. and Macleod,M.D.:Particle filtering applied to musical tempo tracking. Journal of Advances in Signal Processing, vol. 15, pp. 2385–2395 (2004)
10. Ozerov, A., Vincent, E., Bimbot, F.: A General Flexible Framework for the Handling of Prior Information in Audio Source Separation. IEEE Transactions on Audio, Speech, and Language Processing. vol. 19, no. 8. (2011)
11. Marxer, R., Janer, J., Bonada, J.: Low-latency Instrument Separation in Polyphonic Audio Using Timbre Models. In: 10th International Conference on Latent Variable Analysis and Source Separation, LVA/ICA 2012, Tel-aviv, Israel (2012)

Oracle Analysis of Sparse Automatic Music Transcription

Ken O'Hanlon ⋆, Hidehisa Nagano ⋆†, and Mark D. Plumbley⋆ ⋆

⋆Queen Mary University of London
†NTT Communication Science Laboratories, NTT Corporation
{keno,nagano,Mark.Plumbley}@eecs.qmul.ac.uk

Abstract. We have previously proposed a structured sparse approach to piano transcription with promising results recorded on a challenging dataset. The approach taken was measured in terms of both frame-based and onset-based metrics. Close inspection of the results revealed problems in capturing frames displaying low-energy of a given note, for example in sustained notes. Further problems were also noticed in the onset detection, where for many notes seen to be active in the output transcription an onset was not detected. A brief description of the approach is given here, and further analysis of the system is given by considering an oracle transcription, derived from the ground truth piano roll and the given dictionary of spectral template atoms, which gives a clearer indication of the problems which need to be overcome in order to improve the proposed approach.

Keywords: Automatic Music Transcription, Sparse representations

1 Introduction

Automatic Music Transcription (AMT) is the attempt for machine understanding of musical pieces. Many methods proposed for AMT use atomic decompositions of a spectrogram with spectral basis atoms representing musical notes. The atoms may be learned online, using methods such as Non-negative Matrix Factorisation (NMF) [6] or sparse dictionary learning [8]. Alternatively a dictionary may be learnt offline, and the decomposition performed using methods like P-LCA [9] or sparse coding [4].

Often the output from AMT systems is displayed and understood through a piano roll, a pitch time representation relating the onsets and offsets of pitched note events. AMT performance is measured by comparing a computed piano roll with a given ground truth. Often the performance measures are frame-based, with true positives, false negatives and false positives denoted in the derived piano roll and several metrics have been proposed which use these annotations. An alternative perspective to measuring AMT performance is an event-based

⋆ This research is supported by ESPRC Leadership Fellowship EP/G007144/1 and EU FET-Open Project FP7-ICT-225913 "SMALL".

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

analysis [5]. Event-based metrics compare AMT performance in terms of the number of notes for which a correct onset is found within a time-based tolerance.

We have previously proposed an AMT system using structured sparse representations [7] which produced promising results for both frame- and event-based transcription. Visual inspection of the resultant energy-based piano rolls suggests that this approach performs well, capturing much of the energy in the signal, while some limitations are noticed. Often it is found that the energy in the early part of a note is captured, while later sustained elements may be missed, effecting the frame-based analysis. Errors are also noted in the event-based analysis, for which a simple threshold-based onset detection system was used.

These observations lead us to perform an oracle analysis of the system, in order to investigate the causes of these errors, which could possibly reside in either the dictionary used, the transcription system or in the onset detection system. As the system is ultimately based on a (non-negative) least squares analysis, an oracle transcription can be derived by decomposing the signal at each point in time using non-negative least squares (NNLS) with only the atoms representing the notes active, as given by the ground truth, at that time. In the rest of this paper, we describe briefly the AMT system used and the oracle transcription, before analysing the results given by the oracle transcription.

2 Transcription Using Structured Sparse Representations

Sparse representations seek to form the approximation $\mathbf{s} \approx \mathbf{D}\mathbf{t}$ where $\mathbf{s}$ is a signal vector, $\mathbf{D}$ is a dictionary of atoms, and $\mathbf{t}$ is a coefficient vector which is sparse, having few non-zero coefficients. Algorithms for solving sparse representation problems include Orthogonal Matching Pursuit (OMP) [11] which selects, iteratively, the atom most correlated with the residual error and adds this to the support, or collection of selected atoms. At each iteration the supported atoms are backprojected onto the initial signal, giving interim coefficients and a new residual error. Another approach to sparse approximation is the Basis Pursuit (BP) [12], for which many algorithms can be used to solve the optimisation

$$\min_t \|\mathbf{s} - \mathbf{D}\mathbf{t}\|_2^2 + \lambda\|\mathbf{t}\|_1 \tag{1}$$

where the second term is a penalisation term which promotes sparsity.

Music transcription can be thought of as an inherently sparse problem, as only a few notes are active at a given time. In this work non-negative sparse representations are required to decompose the magnitude spectogram. In group or block sparse representations, it is assumed that certain atoms tend to be active together. This assumption can be leveraged for transcription purposes, as in the previous work [7], allowing several atoms to be used together to represent a note, thereby affording the possiblilty to capture better the dynamics of the frequency spectrum of a note, and hopefully reducing the error in the transcription system. In this prior the block of atoms used to represent each note was made of a fixed number, P, of atoms which were adjacent in the dictionary $\mathbf{D} \in \Re^{M \times K}$. Here

$K = L \times P$ where L is the number of groups, thereby defining a set of indices G for the group-based dictionary:

$$G = \{G_l | G_l = \{P \times (l-1) + 1, ..., P \times l\}\} \forall l \in \{1, ..., L\}.$$

In [7] a variant of the Non-negative Basis Pursuit (NN-BP) algorithm [1] was proposed which we call NN-BP(GC). This variant differs from the NN-BP algorithm only through the calculation of a group coefficient, on which the thresholding step is performed, and is outlined in Algorithm 1. Transcriptions using this method had high recall, as many true positives were recovered, while displaying low accuracy as many false positives were also found, though many of the false positives were seen to be of low energy. This poor performance may be due partially to the lack of explicit group penalisation in this method.

A non-negative group version of OMP called Non-negative Nearest Subspace OMP (NN-NS-OMP) was also proposed. This was seen to suffer from a failure to capture low energy atoms, and harmonic jumping was seen to have a negative effect on time continuity in note events in the piano roll. As the method is iterative, a stopping condition needs to be selected, and it was found that selection of an apt stopping condition was tricky.

Algorithm 1 NN-BP(GC)

Input
$\mathbf{D} \in \Re_+^{M \times K}$, $\mathbf{S} \in \Re_+^{M \times N}$, δ, $\mathbf{T}^0 = \mathbf{D}^{\mathrm{T}}\mathbf{S}$, $\Gamma = 1^{L \times N}$
repeat
$t_{k,n} \longleftarrow t_{k,n} \dfrac{[\mathbf{D}^{\mathrm{T}}\mathbf{S}]_{k,n}}{[\mathbf{D}^{\mathrm{T}}\mathbf{D}\mathbf{T}]_{k,n} + \lambda}$
until a fixed number of iterations
$\mathbf{GC}_{l,n} = \sum \mathbf{T}_{G_l,n} \ \forall (l,n)$
$\mathbf{GC}_{l',n'} = 0$; $\Gamma_{l',n'} = 0 \ \ \forall \{l', n'\}$ *s.t.* $\mathbf{GC}_{l',n'} < \delta \times \max \mathbf{GC}$

Molecular sparsity [2] was proposed as an extension of greedy sparse algorithms, in which several atoms related through proximal structure were selected together at each iteration, based on a coefficient system which considered several atoms simultaneously. This approach has the advantage of favouring structure in the decomposition. For example in the Molecular Matching Pursuit (MMP) [2], a molecule of time-persisting tonal elements were extracted from the spectrogram at each iteration by performing tracking through time from an initially selected atom until the onset and offset of the tonal element were found, and all interim atoms were selected.

Initial attempts to build a molecular transcription system were seen to fail when polyphony grew as it became difficult to track pitched atoms (or groups of atoms), due to high projection values being present beyond the onset and offset points of a note, in particular when notes which were similarly pitched or harmonically related were active there. This led to a two-step approach. As previously mentioned the NN-BP(GC) displayed high recall and it was observed

that notes displayed time continuity in otherwise very noisy transcriptions, and it was proposed to first decompose the spectrogram using the NN-BP(GC). Isolated atom supports were pruned and clustering of time-persisting atoms into molecules was performed on the sparse support $\mathbf{\Gamma}$. The molecules were then input to a greedy method called Molecular Non-negative Nearest Subspace OMP (M-NN-NS-OMP) which selects at each iteration one predetermined molecule.

Algorithm 2 M-NN-NS-OMP

Input
$\mathbf{D} \in \Re_+^{M \times K}$, $\mathbf{S} \in \Re_+^{M \times N}$, $\mathit{\Gamma} \in \{0,1\}^{L \times N}$, G, α
Initialise
$i = 0$; $\mathit{\Phi} = 0^{L \times N}$; $B = \{\beta_n | \beta_n = \{\} \forall n \in \{1, .., N\}\}$
repeat
$i = i + 1$
Get group coefs Θ and smoothed coefs $\bar{\Theta}$
$\mathbf{x}_{G_l,n} = \arg\min_x \|\mathbf{r}_n^i - \mathbf{D}_{G_l}\mathbf{x}\|_2^2 \; s.t. \; \mathbf{x} \geq 0 \; \forall l \in \mathit{\Gamma}_n$
$\Theta_{l,n} = \|\mathbf{x}_{G_l,n}\|_1$; $\quad \bar{\Theta}_{l,n} = \sum_{n'=n}^{n+\alpha-1} \Theta_{l,n'}/\alpha$
Select initial atom and grow molecule
$\{\hat{l}, \hat{n}\} = \arg\max_{l,n} \bar{\Theta}_{l,n}$
$n_{min} = \min \bar{n} \; s.t. \; \mathit{\Gamma}_{\hat{l},\Xi} = 1, \; \Xi = \{\bar{n}, ..., \hat{n}\}$
$n_{max} = \max \bar{n} \; s.t. \; \mathit{\Gamma}_{\hat{l},\Xi} = 1, \; \Xi = \{\hat{n}, ..., \bar{n}\}$
$\beta_n = \beta_n \cup \hat{l} \; \forall n \in \Xi = \{n_{min}, ..., n_{max}\}$
Calculate current coefficients and residual
$\mathbf{t}_{G_{\beta_n},n} = \min_t \|\mathbf{s}_n - \mathbf{D}_{G_{\beta_n}}\mathbf{t}\|_2^2 \; \forall n \in \Xi$
$\mathbf{r}_n^{i+1} = \mathbf{s}_n - \mathbf{D}_{G_{\beta_n}}\mathbf{t}_{G_{\beta_n}} \; \forall n \in \Xi$
until stopping condition met

The M-NN-NS-OMP algorithm returns a sparse group coefficient matrix, $\mathbf{T}$, and the transcription performance using this approach was measured with both frame-based and onset-based analysis. The frame-based analysis is performed by comparing a ground truth and the derived transcription. Each frame which is found to be active in both the ground truth and the transcription denotes a *true positive - tp* while frames which are active only in the ground truth and transcription denote *false negatives -fn* and *false positives - fp*, respectively.

For event-based analysis, onset detection was performed on $\mathbf{T}$. An simple threshold-based onset detector was used, based upon the one used in [10] which registered an onset when a threshold value was surpassed and subsequently sustained for a given number of successive frames for a note in the coefficient matrix $\mathbf{T}$. A *tp* was registered when the onset was detected within one time bin of a similarly pitched onset in the ground truth. Similar to the frame-based analysis, an onset found only in the ground truth registered a *fn*, and an onset found only in the transcription registered a *fp*.

Using these markers the following metrics are defined for both frame- and event-based transcription; $Acc = tp \times 100/(tp + fp)$ relates the accuracy of the system in finding correct frames; the recall $Rec = tp \times 100/(tp + fn)$ defines the performance in terms of the amount of correct frames found relative to the number of active frames in the ground truth; $F = 2 * Acc * Rec/(Acc + Rec)$ defines overall performance, considering both false positives and negatives in the measure.

2.1 Experimental Results

Transcription experiments were run using the molecular approach on a set of pieces played on a Disklavier piano from the MAPS [3] database which includes a midi-aligned ground truth. A subdictionary was learnt for each midi note in the range $21 - 108$ from isolated notes also included in the MAPS database, and $\mathbf{D}$ was formed by concatenating these subdictionaries. Transcription was performed using the two-step NN-BP(GC) followed by M-NN-NS-OMP approach.

	Onset-based			Frame-based		
P	Acc	Rec	F	Acc	Rec	F
1	78.3	74.3	76.3	69.1	73.6	71.3
2	78.8	76.2	77.5	69.0	76.4	72.5
3	77.6	77.1	77.4	69.5	78.7	73.8
4	78.8	77.3	78.1	71.8	79.3	75.3
5	78.6	77.8	78.2	72.9	80.0	76.3

Table 1. Frame-based and onset-based transcription results for the proposed molecular approach, relative to the block size, P

We can see from the table of results the performance for both onset-based and frame-based metrics improves with the group size P, thereby validating the use of group sparse representations for this purpose. The experiments were run with a common value used as the stopping condition. Further experiments have shown that improved performance is possible using different values for each group size. In particular, an F-measure greater than 80% was achieved for frame-based transcription for $P = 5$.

3 Transcription Oracle for Sparse Methods

An oracle for transcription performance is proposed. OMP-based methods use a backprojection of the selected atoms onto the signal to produce the final coefficients, thereby gives a (non-negative) least squares error solution with a given support. As the MAPS [3] database comes with a standardised ground truth, we consider an oracle transcription for a given dictionary, given the ground truth

support. At each time bin we calculate the non-negative least squares solution using only the groups of atoms G_n^{oracle}, known from the ground truth to be active at the time bin n.

$$\mathbf{t}_{G_n^{oracle}} = \min_t \|\mathbf{s}_n - \mathbf{D}_{G_n^{oracle}}\mathbf{t}\|_2^2 \; s.t. \mathbf{t} > 0 \; \forall n \in \{1, ...N\} \tag{2}$$

The oracle group coefficient matrix $\mathbf{E}$ is formed by summing the coefficients of the indiviual group members

$$\mathbf{E}_{l,n} = \sum \mathbf{T}_{G_l,n}^{oracle} \; \forall\{l, n\} \tag{3}$$

4 Oracle Analysis

Using this oracle, we can probe the effectiveness of the approach taken to AMT. Interesting observations were made with relation to two aspects of the transcription system; often there is very low energy in supported atoms in $\mathbf{E}$, which may explain how the thresholding in the NN-BP(GC) effected the possible recall rate; secondly, using the oracle transcription provides an insight into the effectiveness of the onset detection system used.

4.1 Energy Based Thresholding

In the NN-BP(GC) algorithm, a thresholding factor δ is used, which is multiplied by the maximum value of the group sparse coefficients $\mathbf{GC}$. For the experiments in [7], a value of $\delta = 0.01$ was used. Using this value for δ it was found that the recall rate of the NN-BP(GC) algorithm in these experiments was 87%, and closer analysis showed that often the false negatives existed at the tail of sustained notes, were it is expected that low energy is displayed. This recovery rate effectively sets an upper bound on the possible recall rate of the M-NN-NS-OMP.

The oracle energy matrix $\mathbf{E}$ was calculated for each piece from the MAPS dataset used in the previous experiments for both ERB and STFT decompositions, both of which used dictionaries learnt from the same dataset of isolated notes in MAPS as used in the previous work [7]. The signals were undersampled to $22.05kHz$, and the ERB spectrogram used 256 frequency bin scale with a $23ms$ time window. The STFT used a 1024 frequency bin spectrogram, with a a 75% overlap, in order to use the same time resolution as the ERB. The NN-BP(GC) was also run for both tranforms to compare the effects of δ thresholding.

The results are displayed in Table 2, where it seen that Rec^{oracle}, the percentage of frames in the oracle transcription $\mathbf{E}$ with higher coefficients than the signal dependent threshold, $th = \delta \times \max \mathbf{E}$ is very similar in both transforms, across all values of delta. A similar pattern is also seen for Rec, the recall rate using the NN-BP(GC), which is smaller than Rec^{oracle}, but again is similar across the transforms, which suggests that the problem here is energy related, and not related to the dictionaries. It can be seen that while the recall rate increases as

	STFT			ERB		
δ	Acc	Rec	Rec^{oracle}	Acc	Rec	Rec^{oracle}
0.1	88.6	38.2	44.1	84.4	37.3	44.6
0.01	38.5	84.6	90.7	36.4	85.0	90.6
0.001	19.2	92.8	96.5	17.7	93.5	96.4
0.0001	12.7	95.2	97.1	12.2	95.6	97.0

Table 2. Analysis of effect of δ on Acc and Rec of NN-BP(GC) and the oracle

δ decreases, the accuracy of the NN-BP(GC) is greatly reduced. Using a smaller value of accuracy might negatively interfere with the final transcription, by introducing oversized molecules and may also effect on the computational load using the current approach as the M-NN-NS-OMP will require more projections.

4.2 Onset Analysis

In the prior work, a simple threshold-based onset detection system was used, which triggered an onset when a threshold value was surpassed and sustained for a minimum length of time. A true positive was flagged when this trigger happened within one time frame of a ground truth onset of the same note. Using the optimal transcription **E** we can test the effectiveness of this onset detection system. Experiments were run using the same parameters as in [7] and the results are presented in Table 3.

P	1	2	3	4	5
Rec	76.2	78.5	79.5	80.1	80.1
Acc	86.4	87.1	87.0	87.3	86.8

Table 3. Onset analysis of oracle transcription **E** for different values of P

The results are not promising given that an oracle transcription is given to the onset detector. Closer inspection of the individual results reveal systematic flaws in the onset detection. False positives are often found when a sustained note is retriggered by oscillation around the threshold value, behaviour which is often found in the presence of other note onsets and may be due to transient signal elements effecting the smoothness of the decomposition across time. Several common types of false negative were found. It is found that a note replayed with minimal time between the offset of the original event and the onset of the following event may produce a false negative where the observed coefficient has not already fallen below the threshold value. When several notes onset simultaneously, onsets may not be detected for all of these notes. A tendency for lower pitched notes not to trigger an onset event in the detection system is also no-

ticed. Further to this we also find some timing errors, where a false negative and a false positive are closely spaced.

5 Conclusion

We have previously proposed an AMT system based on group sparse representations which is relatively fast and shows promising results. An oracle transcription has been presented here, which gives some insight into the some weaknesses in the AMT system, as currently exists. Further work will focus on improving the AMT system, by incorporating a more sophisticated onset detection system and possibly using a new algorithm to perform the decomposition.

References

1. Aharon, M., Elad, M., Bruckstein, A. M.: K-SVD and its non-negative variant for dictionary design. In: Proc. of the SPIE conference wavelets, 2005, pp. 327-339
2. Daudet, L.: Sparse and structured decompositions of signals with the molecular matching pursuit. In: IEEE Transactions on Audio, Speech and Language Processing, 2006, pp. 1808-1816
3. Emiya, V., Badeau, R., David, B.: Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. In: IEEE Transactions on Audio, Speech and Language, 2010, pp. 1643-1654
4. Leveau, P., Vincent, E., Richard, G., Daudet, L.: Instrument-Specific Harmonic Atoms for Mid-Level Music Representation. In: IEEE Transactions on Audio, Speech and Language, 2008, pp. 116-128
5. Poliner, G., Ellis, D.: A discrimative model for polyphonic piano transcription. In: EURASIP Journal Advances in Signal Processing, no. 8, 2007, pp. 154-162
6. Smaragdis, P., Brown, J. C.: Non-negative matrix factorization for polyphonic music transcription. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2003
7. O'Hanlon, K., Nagano, H., Plumbley, M. D.: Structured Sparsity for Automatic Music Transcription. In: IEEE Int. Conference on Audio, Speech and Signal Processing 2012.
8. Abdallah, S.A., Plumbley, M. D.: Polyphonic transcription by non-negative sparse coding of power spectra. In: Proceedings ISMIR 2004, pp. 318-325
9. Benetos, E., Dixon, S.: Multiple-Instrument polyphonic music transcription using a convolutive probabilistic model. In: Proceedings of the Sound and Music Computing Conference 2011
10. Bertin, N., Badeau, R., Vincent, E.: Enforcing harmonicity and smoothness in bayesian non-negative matrix factorization applied to polyphonic music transcription. In: IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 3, pp. 538549, Mar 2010.
11. Pati, Y. C., Rezaiifar, R.: Orthogonal Matching Pursuit: Recursive function approximation with applications to wavelet decomposition. In: Proceedings of the 27th Annual Asilomar Conference on Signals, Systems and Computers, 1993, pp. 40-44.
12. Chen, S. S., Donoho, D. L., Saunders, M. A.: Atomic decomposition by Basis Pursuit. In: SIAM Journal on Scientific Computing, vol. 20, pp. 33-61, 1998.

Oral session 6:

Film Soundtrack and Music Recommendation

The Influence of Music on the Emotional Interpretation of Visual Contexts

Designing Interactive Multimedia Tools for Psychological Research

Fernando Bravo

University of Cambridge. Centre for Music and Science.
nanobravo@fulbrightmail.org

Abstract.

From a cognitive standpoint, the analysis of music in audiovisual contexts presents a helpful field in which to explore the links between musical structure and emotional response.

This work emerges from an empirical study that shows strong evidence in support of the effect of tonal dissonance level on interpretations regarding the emotional content of visual information.

From this starting point it progresses toward the design of interactive multimedia tools aimed at investigating the various ways in which music may shape the semantic processing of visual contexts. A pilot experiment (work in progress) using these tools to study the emotional effects of sensory dissonance is briefly described.

Keywords: Music, emotions, film-music, interactive multimedia, algorithmic composition, dissonance, tonal tension, interval vector, Max/MSP/Jitter.

1 Introduction and Background

Although research in music cognition has been growing steadily during the past four decades, we still lack a significant body of empirical studies concerning the higher levels of musical response, including the emotional and aesthetic aspects. From a cognitive standpoint, the analysis of music in audiovisual contexts presents a helpful field in which to explore the affective and connotative aspects of musical information [1, 2, 3].

This paper describes work in progress to investigate the influence of tonal dissonance on the emotional interpretation of visual information.

The objectives of this paper are:

- To report on a formal experiment showing the effect of tonal dissonance on interpretations regarding the emotional content of an animated short film (Section 2).
- To describe a series of interactive multimedia tools designed to investigate the various ways in which music may shape the semantic processing of visual contexts (Section 3).

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

- To show an example of how these tools could be used in experimental cognition research. In this example, I employ stochastically generated music to empirically study the links between sensory dissonance and emotional responses to music in a strictly controlled audiovisual setting (Section 3.2).

Consonance and dissonance refer to specific qualities an interval can posses [4]. Tonal and sensory dissonance are sometimes used as equivalent concepts. However, as Krumhansl [5] has expressed, these two notions have different shades of meaning. Sensory dissonance designates, first of all, a psychoacoustic sensory property associated with the presence/absence of interaction between the harmonic spectra of two pitches [6]. Tonal dissonance includes sensory dissonance but it also captures a more cognitive or conceptual meaning beyond psychoacoustic effects that is typically expressed with terms such as tension or instability. The term "tonal dissonance", as employed here, refers both to sensory and cognitive dissonance.

Meyer proposed a theory of meaning and emotion in music [7]. According to his assumptions the confirmation, violation or suspension of musical expectations elicits emotions in the listener. Following this theory, researchers found association between specific musical structures, precise neural mechanisms and certain neurophysiological reactions that are strongly connected with emotions. In addition, studies focusing on the perception of tonal dissonance have shown that unexpected chords and increments in dissonance have strong effects on perceived tension [5, 8, 9, 10], which has been linked to emotional experience during music listening [11].

Tonal dissonance can be described by a number of variables [9], which have been already historically studied by music theorists and scientists: the tonal function of chords inside a musical context [12, 13, 14, 15, 16], their acoustic or sensory consonance [6, 17], and melodic organization, usually referred as "horizontal motion" [18].

Cognitive approaches usually emphasize the importance of melodic organization and tonal function while sensory-perceptual theories tend to focus on psychoacoustical aspects. In this paper, I use the term 'tonal dissonance' as a synonym for 'tonal tension', to refer to the effects of tonal function, sensory dissonance and horizontal motion on perceived musical tension.

2 Experimental Investigation

This paper emerges from a formal experiment entitled "The influence of tonal dissonance on emotional responses to film" [19]. The main experimental hypothesis predicted that, within the same film sequence (visual context), different musical settings, in terms of tonal dissonance, would systematically elicit different interpretations and expectations about the emotional content of the same movie scene.

2.1 Experimental Design

This experiment was aimed at addressing the particular emotional effect of tonal dissonance induced by chord changes, controlling for other elements within musical

structure such as tempo, intensity, rhythm, timbre (instrumentation), etc. This was achieved by working with a precise experimental design, also used by Blood *et al.* in their neuroscientific research (which investigated the cerebral activations elicited by tonal dissonance) [20]. It is important to note that this study sets aside other kinds of musical tension. Empirical evidence has shown that musical tension can be induced by many factors, such as rhythm, dynamics, tempo, gesture, textural density and tone timbre [25, 26, 27]. This work focuses on musical tension induced by tonal dissonance in the specific sense of tension created by melodic and harmonic motion.

A choral piece, specifically composed for the experiment, was made to sound more or less consonant or dissonant by modifying its harmonic structure, producing two otherwise-identical versions of the same music passage. These two contrasting conditions, in terms of tonal dissonance, were used as background music for the same passage of an animated short film ("Man with pendulous arms" - 1997, directed by Laurent Gorgiard).

Table 1. Cross-classification of music condition and response variable (number of participants and percentage of participants within condition)

The character	*feels confident*		*is scared*	
consonant	37	61.7%	23	38.3%
dissonant	25	41.7%	35	58.3%
The mood of the story is	*nostalgic*		*sinister*	
consonant	58	96.7%	2	3.3%
dissonant	26	43.3%	34	56.7%
The character is trying	*to create something*		*to destroy something*	
consonant	45	75%	15	25%
dissonant	27	45%	33	55%
The character	*is a fantasy character*		*is monstrous*	
consonant	53	88.3%	7	11.7%
dissonant	39	65%	21	35%
Genre of the short film	*Drama*		*Horror*	
consonant	59	98.3%	1	1.7%
dissonant	42	70%	18	30%
The character	*is alienated*		*is sad*	
consonant	11	18.3%	49	81.7%
dissonant	35	58.3%	25	41.7%
Character's actions	*directed by his own will*		*external influence*	
consonant	50	83.3%	10	16.7%
dissonant	35	58.3%	25	41.7%
The end of the short film	*will probably be hopeful*		*will probably be tragic*	
consonant	41	68.3%	19	31.7%
dissonant	29	48.3%	31	51.7%
The character is trying	*to protect himself*		*to search something*	
consonant	47	78.3%	13	21.7%
dissonant	44	73.3%	16	26.7%

A total of 120 healthy volunteers with normal hearing took part in this experiment. The participants were randomly sampled from students at Argentine Catholic University. Two independent samples were used (60 participants each). The subjects were randomly assigned to two groups, one of which saw an animated short film with the "consonant music" condition and the other saw the same film with the "dissonant music" condition. At the end participants were asked to answer a survey about their associations and expectations towards the main character and the overall story of the film. The survey used 9 single-selection questions, asking participants to choose only one item from two items given. Table 1 shows participants' answers within each music condition.

2.2 Experimental Results

Eight out of nine response variables were found associated with the explanatory variable (tonal dissonance level). The variable related to the character's objective (at the bottom of Table 1) was the only variable that did not reach significant association with tonal dissonance level.

For the eight response variables where association was found, two ways to summarize the strength of the association are presented: the *difference of proportions*, forming confidence intervals to measure the strength of the association in the population, and the *odds ratio* (Table 2).

When measuring the strength of the association, variables related to the mood in the story, the emotional state of the character and the interpreted genre of the short film were found to have the strongest association with dissonance level in background music (grey cells).

Table 2. $\chi 2$, Difference of proportions and Odds Ratio

		Differ. of proportions		Odds Ratio	Odds	
Variable	$\chi 2$(*p value*)	$p1$-$p2$	95% CI	OR	Con	Dis
Intentions (create/destroy)	11.2(<.01)	0.3	[.133, .467]	3.667	3	0.8
Feeling (confident/scared)	4.80(<.05)	0.2	[.025, .035]	2.252	1.6	0.7
Mood (nostalgic/sinister)	40.6(<.01)	0.53	[.401, .667]	37.92	29	0.7
Emot.state (sad/alienated)	20.3(<.01)	0.4	[.241, .559]	6.236	4.4	0.7
Actions (own will/external)	9.07(<.01)	0.25	[.094, .406]	3.571	5	1.4
Class (fantasy/monstruous)	9.13(<.01)	0.23	[.087, .379]	4.077	7.5	1.8
Genre (drama/horror)	18.0(<.01)	0.28	[.163, .403]	25.28	59	2.3
Ending (hopeful/tragic)	4.93(<.05)	0.2	[.027, .373]	2.307	2.1	0.9

For example, Table 2 shows that there was a rise of 0.4 in the proportion that interpreted the emotional state of the character as sad among participants who saw the film with consonant music. Also, we may infer, with 95% confidence, that $p1$ (the proportion of people seeing the film with consonant music and interpreting the character's

emotional state as sad) may be as much as between [0.241, 0.559] larger than $p2$ (the proportion of people seeing the film with dissonant music and interpreting the character's emotional state as sad).

In addition, from Tables 1 and 2, we observe that for the consonant music condition the proportion of people who interpreted the character's emotional state as sad equals 49 / 11 = 4.4545. The value of 4.45 means that, for participants who saw the film with consonant music, there were 4.45 participants who interpreted the character's emotional state as sad, for every 1 person in the dissonant condition. On the other hand, for the dissonant music condition the proportion of people who interpreted the character's emotional state as sad equals 25 / 35 = 0.7143. Equivalently, since 35 / 25 = 1 / 0.7143 = 1.4, this means that there were 1.4 participants in dissonant condition who interpreted the character's emotional state as alienated for every 1 person in consonant condition. For the consonant music condition, the odds of interpreting the character's emotional state as sad were about 6.2 times the odds of the same interpretation for the dissonant music condition.

The results of this experiment offer strong evidence in support of the effect of tonal dissonance level (in film music) on interpretations regarding the emotional content of visual information.

2.3 Discussion of Experimental Results

The empirical research described supports and confirms previous research on mood congruency effects [1], and can be interpreted within Annabel Cohen's Congruence-Associationist framework of the mental representation of multimedia [2, 3].

In this work, tonal dissonance level was experimentally isolated in order to analyze a particular feature within the multiple musical structures that may elicit musical emotions. As pointed in section 2.1., this study was focused on musical tension induced by chord changes. Other important factors that contribute to the building and release of musical tension, such as timbre, dynamics, textural density, etc., which were controlled in the experiment, are not examined in the present discussion.

Results revealed that the background music significantly biased the affective impact of the short film. Generally, the consonant music condition guided participants toward positive emotional judgments, while dissonant music guided participants toward negative judgments. In addition, the dissonant background music seems to have rendered the interpretation more ambiguous when compared to the higher percentages for the positive judgments in the consonant condition. However, additional research is needed to further examine this hypothesis since the present experiment did not include a visual alone condition, which would be necessary to control for the effects of visual content by itself.

Music theory provides technical descriptions of how styles organize musical sounds and offers insights about musical structures that might underlie listeners' interpretations. Within the general perspective of post-tonal music theory, Allen Forte has introduced the notion of interval-class content [21]. This concept, widely used in the analysis of atonal twentieth-century music, offers an interesting approach to qualifying sonorities. A pitch interval is simply the distance between two pitches, meas-

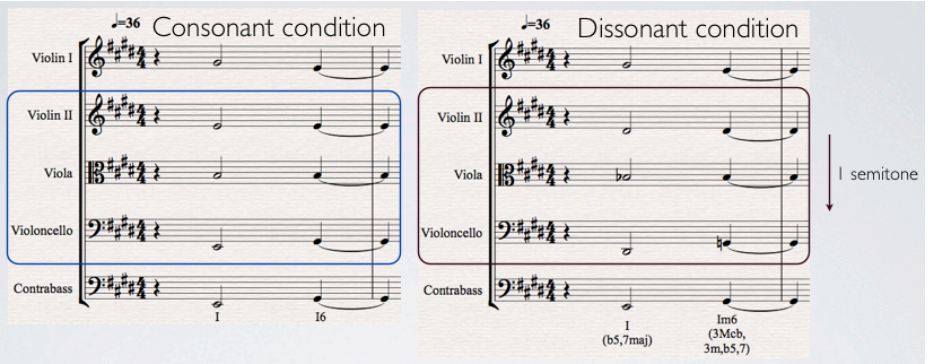
ured by the number of semitones. The ordered pitch intervals (ascending or descending) focus attention on the contour of the line. The unordered pitch intervals ignore direction of motion and concentrate entirely on the spaces between the pitches. An unordered pitch-class interval is the distance between two pitch classes, and it is also called interval class [28]. Because of octave equivalence, compound intervals (intervals larger than an octave) are considered equivalent to their complements in mod 12. In addition, pitch-class intervals larger than six are considered equivalent to their complements in mod 12. The number of interval classes a sonority contains depends on the number of distinct pitch classes in the sonority. For any given sonority, we can summarize the interval content in scoreboard fashion by indicating, in the appropriate column, the number of occurrences of each of the six interval classes (occurrences of interval class 0, which will always be equal to the number of pitch classes in the sonority, are not included). Such scoreboard conveys the essential sound of a sonority.

Table 3 summarizes interval class content for the first measure of the experimental transformation used in this study to create contrasting conditions (Figure 1). The comparative dissonant condition was obtained by lowering, by a semitone, the second violin, viola and violoncello lines, while keeping the other instruments in their original position (at their original pitch). Thus, the level of dissonance was uniform throughout a given version. The analysis, therefore, can generally represent the comparative level of dissonance throughout.

Table 3. Interval Content of the two music conditions

Consonant condition - Interval Class content						
Interval Class	1	2	3	4	5	6
No. of occurrences	0	0	3	6	5	0
Dissonant condition - Interval Class content						
Interval Class	1	2	3	4	5	6
No. of occurrences	4	2	2	3	6	2

Fig. 1. Score of the consonant and dissonant music conditions (first measure)



The consonant condition is primarily governed by collections of intervals considered to be consonant (thirds, fourths and fifths). In contrast, the use of dissonant intervals in the dissonant version (major second, minor second and tritone) has a very specific emotional effect that is reflected in the participants' interpretations.

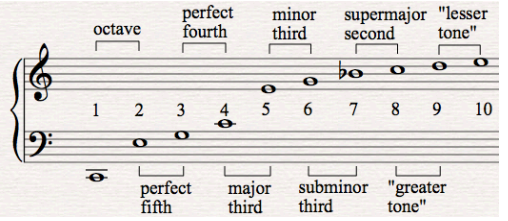
3 Future Work: Interactive Multimedia Tools for Experimental Research on Interval Content and Musical Emotions

The described experiment stimulated the investigation of tonal [15] and post-tonal [21] interval theory, and the parallel design of interactive multimedia tools to empirically analyze the effects of interval content on musical emotions.

3.1 Background Elements

Within the tonal perspective, Paul Hindemith's work is especially noteworthy [15]. According to Hindemith, the overtone series system (see Figure 2) gives a complete proof of the natural basis of tonal relations. In general, as new intervals are introduced, the stability decreases and the two tones involved are considered more distant in their relation. All music theories have a general agreement on this model.

Fig. 2. Overtone series with intervals labeled



This theory has several links with the concept of sensory dissonance as studied in the psycho-acoustic literature [6, 22, 23, 24]. According to this model, the most consonant intervals would be the ones that could be expressed with simple frequency ratios, which has been supported by psychological study. Intervals such as the unison (1:1), the octave (2:1), perfect fifth (3:2), and perfect fourth (4:3) are regarded as the most consonant. Intermediate in consonance are the major third (5:4), minor third (6:5), major sixth (5:3), and minor sixth (8:5). The most acoustically dissonant intervals (composed of frequencies the ratio between which is not simple) are the major second (9:8), minor second (16:15), major seventh (15:8), minor seventh (16:9), and the tritone (45:32).

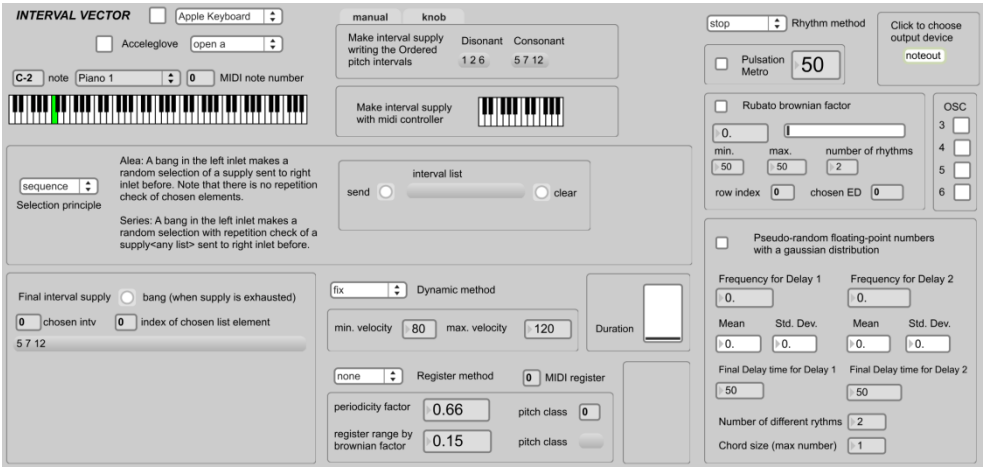
From the perspective of atonal theory, Allen Forte's work provided a general theoretical framework from where to start the exploration of intervals in a new way, a way that was intimately concerned with the idea of sonority [21]. He explained that different types of sonorities could be generally defined by listing their constituent intervals. In the previous experiment, I showed how the two music conditions could be de-

scribed in terms of interval content. Forte introduced the basic concept of "interval vector" to analyze the properties of pitch class sets and the interactions of the components of a set in terms of intervals. An interval vector is an array that expresses the intervallic content of a set. It has six digits, with each digit standing for the number of times an interval class appears in the set [21]. According to Forte, such interval vector conveys the essential sound (color, quality) of a sonority.

3.2 'Intermedia Patch'. Cross-modal Research on Intervals and Visuals

The interactive multimedia tools presented in this section, called 'Intermedia patch', were built to explore the interval vector theory in a practical and strictly controlled setting, in order to experimentally study the links between sonority and emotional response. The patch works with an initial supply of intervals and allows to experiment with different algorithmic composition techniques, allowing a detailed control over many coincident variables such as loudness, rhythm, timbre, melody, intensity and instrumentation.

Fig. 3. Intermedia patch built with Cycling'74 Max/MSP/Jitter (top) that allows to simultaneously work with images created with Maxon Cinema 4D software (bottom)



The patch not only provides a programming environment for analyzing different types of sonorities based on interval selection (Figure 3 top), it also allows to simultaneously work with images (Figure 3 bottom), enabling the study of mood congruency effects between sound and visuals.

The tool is currently being tested in a pilot study (in progress), which employs the patch for the creation of sound stimuli. In this experiment I opted to analyze the emotional reactions induced by interval content. Participants are asked to see a short animation created for this study, with stochastically generated background music.

Participants are randomly assigned to three independent groups; one control group sees the animation without music, a second group sees the animation with a consonant interval content as background music (interval set: 5-7-12, all perfect consonances), and a third group sees the same animation with a dissonant interval content (1-2-6, all dissonances). Immediately after viewing the clip, participants are asked to complete a series of bipolar adjective ratings representing the three connotative dimensions: activity, potency and valence.

The question posed in this study is whether two contrasting examples of background music, in terms of interval content, can selectively bias observers' emotional interpretation of visual information. People who have internalized the Western tonal music conventions normally respond to certain sonorities in a specific manner. The main experimental hypothesis predicts that, in particular, the valence dimension should differ significantly under these two conditions. Positive results would confirm mood congruency effects induced exclusively by interval content (surface or sensory consonance).

4 Conclusions

The empirical research included in this paper supports and confirms previous studies that have examined, from a cognitive perspective, the role of music on the interpretation of a film or a video presentation [1, 2, 3]. The results offer strong evidence in support of the effect of tonal dissonance level on interpretations regarding the emotional content of visual information. Moreover, it gives insights to the richness and potentiality of the aural "palette", since extensive effects on the emotional interpretation of visual contexts may be directed by the manipulation of a single musical structure feature (tonal dissonance).

Studies such as this demonstrate associations between aspects of musical structure and musical meaning, which then becomes automatically attached to the visual content or implied narrative that is in the focus of the spectator's attention.

The positive results of this study indicate that further research that systematically examines the multiple and subtle ways in which music performs elaborative functions in the comprehension of visual contexts should be pursued. The interactive multimedia tools introduced in section 3 are aimed at exploring this path. These tools incorporate a variety of potential variables in both musical sound and transformations of the visual stimuli for experimental purposes, providing a foundation on which future research could build.

Acknowledgments. Thanks to Prof. Ian Cross, Prof. Sarah Hawkins and to all the researchers at the Centre for Music and Science (University of Cambridge). Thanks to Dr. Christopher Hopkins, Prof. Anson Call and Prof. Steve Herrnstadt for their constant support. Thank you to the anonymous reviewers for their suggestions that improved the paper considerably. This work was conducted at the University of Cambridge and is supported by a Queens' College Walker Studentship.

References

1. Boltz, M. G. (2001). Musical Soundtracks as a Schematic Influence on the Cognitive Processing of Filmed Events. *Music Perception,* 18, 427-454.
2. Cohen, A. J. (2001). Music as a source of emotion in film. In Juslin P. & Sloboda, J. (Eds.). Music and emotion. (pp.249-272). Oxford: Oxford University Press.
3. Cohen, A. J. (2005). How music influences the interpretation of film and Video: Approaches from experimental psychology. In R.A. Kendall & R. W. Savage (Eds.). Selected Reports in Ethnomusicology: Perspectives in Systematic Musicology, 12, 15-36.
4. Bharucha, J.J. (1984). Anchoring effects in music: the resolution of dissonance. *Cognitive Psychology*, 16, 485-518.
5. Lerdahl, F., & Krumhansl, C. L. (2007). Modeling tonal tension. *Music Perception*, 24, 329-366.
6. Helmholtz, H. von (1954). *On the Sensation of Tone as a Physiological Basis for the Theory of Music*. New York: Dover. (Original German work published 1863).
7. Meyer, L. B. (1956). *Emotion and meaning in music*. Chicago: University of Chicago Press.
8. Bigand, E., Parncutt, R., & Lerdahl, F. (1996). Perception of musical tension in short chord sequences: The influence of harmonic function, sensory dissonance, horizontal motion, and musical training. *Perception & Psychophysics*, 58, 124-141.
9. Bigand, E., & Parncutt, R. (1999). Perception of musical tension in long chord sequences. *Psychological Research*. 62, 237-254.
10. Krumhansl, C.L. (1996). A perceptual analysis of Mozart's Piano Sonata, K. 282: Segmentation, tension and musical ideas. *Music Perception*, 13, 401-432.
11. Steinbeis, N., Koelsch, S., & Sloboda, J. A. (2006). The role of harmonic expectancy violations in musical emotions: Evidence from subjective, physiological, and neural responses. *Journal of Cognitive Neuroscience*, 18(8), 1380-1393.
12. Riemann, H. (1896). *Harmony simplified* (H. Bewerung, Trans.). London: Augener. (Original work published 1893).
13. Koechlin, C. (1930). *Traité de l'harmonie*. Paris: Max Eschig.
14. Schenker, H. (1979). *Free Composition* (E. Oster, Trans.). New York: Longman. (Original work published 1935).
15. Hindemith, P. (1942). *Unterweisung im Tonsatz*, 3 vols. (Mainz: Schott, 1937–70) [English edition, as *The Craft of Musical Composition*, vol. 1: Theoretical Part, trans. by Arthur Mendel (New York: Associated Music Publishers; London: Schott, 1942)].
16. Costère, E. (1954). *Lois et styles des harmonies musicales*. Paris: Presses Universitaires de France.
17. Rameau, J.P. (1971). *Treatise of Harmony* (P. Gosset, Trans.). New York: Dover. (Original work published 1722).

18. Ansermet, E. (1961). *Les fondements de la musique dans la conscience humaine*. Neuchâtel: Delachaux et Niestle.
19. Bravo, F. (2011). *The influence of music on the emotional interpretation of visual contexts* by *Bravo, Fernando.* Master's Thesis*,* Iowa State University, United States. AAT 1494771.
20. Blood, A.J., R.J. Zatorre, P. Bermudez & A.C. Evans. (1999). Emotional responses to pleasant and unpleasant music correlate with activity in paralimbic brain regions. *Nat. Neurosci*. 2: 382–387.
21. Forte, A. (1973). *The Structure of Atonal Music*. Yale University Press.
22. Plomp, R. & Levelt, W.J.M. (1965). Tonal consonance and the critical bandwidth. *Journal of the Acoustical Society of America, 38*, 548-560.
23. Vos, J. & van Vianen, B.G. (1984). Thresholds for discrimination between pure and tempered intervals: The relevance of nearly coinciding harmonics. *Journal of the Acoustical Society of America, 77*, 176-187.
24. DeWitt, L.A. & Crowder, R.G. (1987). Tonal fusion of consonant musical intervals. *Perception & Psychophysics, 41*, 73-84.
25. Barthet, M., Depalle, P., Kronland-Martinet, R., Ystad, S., (2010). From clarinet control to timbre perception. *Acta Acustica united with Acustica*, 96, 678-689.
26. Barthet, M., Depalle, P., Kronland-Martinet, R., Ystad, S., (2010). Acoustical correlates of timbre and expressiveness in clarinet performance. *Music Perception*, 28, 135-153.
27. Paraskeva, S., McAdams, S. (1997). Influence of timbre, presence/absence of tonal hierarchy and musical training on the perception of tension/relaxation schemas of musical phrases. *Proceedings of the 1997 International Computer Music Conference, Thessaloniki*, pp.438-441.
28. Strauss, J.N. (1990). *Introduction to Post-Tonal Theory*. Prentice-Hall, Cliffs.

The Perception of Auditory-visual Looming in Film

Sonia Wilkie and Tony Stockman

Queen Mary University of London

sonia.wilkie@eecs.qmul.ac.uk
tony.stockman@eecs.qmul.ac.uk

Abstract. Auditory-visual looming (the presentation of objects moving in depth towards the viewer) is a technique used in film (particularly those in 3D) to assist in drawing the viewer into the created world. The capacity of a viewer to perceptually immerse within the multidimensional world and interact with moving objects, can be affected by the sounds (audio cues) that accompany these looming objects. However the extent to which sound parameters should be manipulated remains unclear. For example, the amplitude, spectral components, reverb and spatialisation can all be altered, but the degree of their alteration and the resulting perception generated, need greater investigation. Building on a previous study analysing the physical properties of the sounds, we analyse peoples responses to the complex sounds which use multiple audio cues for film looming scenes, reporting which conditions elicited a faster response to contact time, causing the greatest amount of underestimation.

Keywords: Auditory-Visual Looming; Sound Design; Psychoacoustics

1 Introduction

A feature of film and gaming is interacting with objects that move in space, particularly objects that move in depth towards the viewer. Examples can be seen in 3-D presentations where objects appear to leap out of the screen towards the viewer; and in gaming where judgements are made to avoid or attack approaching objects.

The sound that accompanies these looming objects can affect the extent to which a viewer can perceptually immerse within the multidimensional world and interact with the moving objects. To accurately generate a dynamic and rich perception of the looming objects, the design of such complex sounds should be based on a firm scientific foundation that encompass' what we know about how we visually and aurally perceive events and interactions.

2 Previous Research and Practice

Previous research on auditory looming has revealed that people associate an approaching object with at least three attributes of sound, including interaural temporal differences, frequency change, and amplitude change [1].

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

In addition to finding that all three attributes of sound were associated with a looming object, they found that the change in amplitude elicited the fastest response to contact time, at the point in which the object passed, whilst the change in frequency prompted a response before the object had passed [1]. This underestimation of the contact time of a looming object, implies that the object is approaching at a faster rate and is anticipated to contact sooner.

Later studies on auditory looming showed that people overestimate the magnitude of intensity when presented with increasing stimuli [2],[3]. This implies that the increasing intensity of the approaching object is more dramatic than it physically is.

In an evolutionary context for both the physical and virtual worlds, these overestimations of magnitude and underestimation of contact time provide an advantage to the observer, giving them more time to prepare (an increased safety margin) for the objects arrival, and to initiate the appropriate response (being fight or flight), therefore increasing the chance of survival.

However, many of these previous auditory looming perception experiments [1],[2],[3],[4], have been conducted in extremely controlled conditions, with the aural stimuli consisting of simple tones (often a sine or triangle wave at 400 - 1000 Hz), and sound parameter manipulations such as an amplitude increase (between 10 - 30 dB), frequency change (using 804 Hz - 764.6 Hz, and 602.9 Hz - 572 Hz, which in musical terms equates to the tone and deviation of G5 $\pm$ 43 cents, and D5 $\pm$ 45 cents), and interaural temporal differences (a delay between the channels from 0.557 ms to 0.00 ms).

Limiting these variables used in experimental conditions compromises the ecological validity of the results, sound parameters manipulated, and real world application.

In contrast however, the film and gaming industries require sound designers to manipulate complex sounds, with the purpose of maximising the viewers experience, immersiveness, responsiveness to onscreen action, and overall perception of the virtual environment.

Examination of the sound manipulation techniques that sound designers and post production technicians use as cues for an approaching object in looming scenes provides a basis for a broader range of variables that can then be used in psychological studies on the perception of approaching objects.

Building upon our previous research [5] that examined the audio cues and techniques that sound designers use to generate the perception of an object moving in depth (looming), this research examines the percepts generated by the complex sounds.

3 Feature Analysis Studies

DSP analysis was previously conducted on the audio track of the 27 film looming scene samples used in this study, to understand which features the sound designers and post production technicians were using as cues for auditory looming,

how the features were manipulated, and the degree of the manipulation. Features that were analysed include: amplitude change; amplitude levels; amplitude slope; interaural amplitude differences; pan position; spectral centroid; spectral range; spectral spread; spectral flux; reverb; roll-off; and image motion tracking of the object.

In summary, our findings showed a number of similar techniques existed between the variety of samples. This includes:

- An average amplitude increase of 62.68 dB ($SD = 15.49$) on a linear / near-linear slope.
- The pan position centrally placed, and close to the image position, however fluctuates more than the image position. This fluctuation emphasises the spatial movement without having to hard pan to a single channel.
- An average spectral centroid increase of 1673.36 Hz.
- An average spectral flux increase of 167.0 Hz (with an average amount of flux of 13.8 Hz at the start of the sample, and 180.8 Hz at the peak).

In contrast to the previous auditory looming studies, the feature analysis of the film samples showed that they have:

- A greater range of variables used simultaneously to form complex looming stimuli (compared to the simple waves in the psychoacoustic studies).
- A greater increase in the levels that the variables were manipulated (ie 62.68 dB amplitude increase in the film samples, versus 10 - 30 dB in the psychoacoustic studies).

4 An Investigation of Responses to Complex Looming Sounds

This study is an extension of our previous research which examined the sound features in the looming samples, and will examine subjects responses to the looming stimuli that uses complex sounds produced by the sound designers and technicians.

4.1 Aim

The aim of this study is to determine if a subjects response to a looming object differs with the inclusion of complex designed sounds that use multiple audio cues, as opposed to looming scenes with no sound.

4.2 Hypothesis

It is hypothesised that the combination of the multimodal (auditory-visual) presentation (with the greater number of cues used, and the greater amount of stimuli change) will cause people to underestimate the contact time of the approaching object, thereby eliciting a faster response time than the looming scenes with no sound.

4.3 Method

4.3.1 Participants

A sample of 15 participants naive to the study purpose were recruited. They were Ph.D students and Postdoc. researchers from Queen Mary, University of London aged between 20 and 36 years ($\mu = 27.07$ years, $SD = 4.70$), with more male participants than female participants (11 male, 4 female).

4.3.2 Stimuli

The stimuli consisted of 27 film scenes that presented objects moving towards the viewer, and were comprised of both auditory and visual components. The scenes used are listed in *Table 1*. They were presented via computer with the visual stimulus presented on the monitor, and the auditory stimulus output through a pair of headphones.

The 27 scenes were presented in each of the three conditions - the multimodal (sound and image) condition, and the two unimodal conditions (sound only or image only). Each trial condition was presented once only (totaling 81 trial presentations) and in a randomised order.

4.3.3 Apparatus

Participants were located at a computer workstation with their head distanced approximately 40 cm from the computer monitor and eyes level with the centre of the monitor.

A Mac Pro 1.1 with a NEC MultiSync EA221WM (LCD) monitor was used. The screen size was 22 inches with the resolution set to 1680 x 1050 pixels and the display was calibrated to a refresh rate of 60 Hz.

The auditory stimulus was presented through Sennheiser HD515 headphones.

The program MAX / MSP / Jitter version 4.6 was used to construct the software application that presented the auditory and visual stimuli; presented the trials in a randomised and collected order, timed the participants responses using the computer's internal clock, and collected the participant responses in a text file.

4.3.4 Procedure

Participants sat at the computer workstation and were informed of the experiment procedure. They were given an information sheet summarising both the procedure and the ethics approval, signed a consent form, and completed a background questionnaire asking questions on gender, age, cinema experience and whether they have had corrections made to their vision or hearing.

Before commencing the experiment, the participants completed a practise test using 6 looming scenes (that were not additionally presented in the experiment). It was conducted as a supervised learning procedure to provide them with the opportunity to comprehend the experiment, the procedure, the micro time scale of the stimulus, and how to complete the task.

Participants were then instructed to start the experiment when ready.

#	Title	Year	Chapter, Time (min : sec)
1	The Matrix	1999	Chapter 1, 1:22 - 1:25
2	Star Wars (*Return of the Jedi*)	1983	Chapter 3, 0:20 - 0:24
3	Star Wars (*Revenge of the Sith*)	2005	Chapter 31, 3:08 - 3:09
4	X-men (*The Last Stand*)	2006	Chapter 15, 0:35 - 0:36
5	The Day After Tomorrow	2004	Chapter 12, 2:29 - 2:33
6	King Arthur	2004	Chapter 7, 10:46 - 10:48
7	Sherlock Holmes	2009	Chapter 22, 4:36 - 4:38
8	Van Helsing	2004	Chapter 17, 1:52 - 1:54
9	I Am Legend	2007	Chapter 17, 0:00 - 0:03
10	Troy	2007	Chapter 27, 2:22 - 2:24
11	Beowulf	2007	Chapter 2, 4:03 - 4:05
12	The Bourne Identity	2002	Chapter 12, 2:10 - 2:12
13	Charlie & the Chocolate Factory	2005	Chapter 15, 1:24 - 1:26
14	Mr and Mrs Smith	2005	Chapter 20, 0:40 - 0:44
15	Sin City	2005	Chapter 18, 1:06 - 1:07
16	28 Days Later	2002	Chapter 11, 0:01 - 0:04
17	Gattaca	1997	Chapter 21, 2:39 - 2:40
18	Alice in Wonderland	2010	Chapter 15, 0:19 - 0:20
19	Avatar	2009	Chapter 22, 1:42 - 1:45
20	Clash of the Titans	2010	Chapter 13, 4:11 - 4:13
21	Despicable Me	2010	Chapter 18, 2:23 - 2:24
22	Kill Bill vol2	2004	Chapter 6, 0:03 - 0:06
23	Mission Impossible 3	2006	Chapter 4, 1:06 - 1:08
24	Yogi Bear	2010	Chapter 1, 1:25 - 1:27
25	Final Destination	2009	Chapter 15, 0:06 - 0:07
26	Salt	2010	Chapter 9, 3:13 - 3:14
27	Saving Private Ryan	1998	Chapter 19, 3:17 - 3:21

Table 1. List of film scenes that were used in the experiment.

The task required participants to watch and/or listen to the scene of an approaching object, and to press the keyboard 'space bar' when they thought the object was closest to them.

Each trial lasted for a total duration of 0.5 - 4.0 seconds (depending on the looming scene presented) and a 6 second break was given between each trial. With a total of 81 trial presentations, the experiment lasted for approximately 25 minutes.

Participants were not given any information implying there might be correct, incorrect or preferred responses.

4.4 Results

Image motion tracking was previously performed on each scene to determine the approaching objects position and size, over time. For the purpose of this

study, the time (of the frame) in which the object was largest was considered the contact point and is called the 'peak'.

Participants responses to the stimuli (by pressing the keyboard 'space bar' when they thought the object was closest) was timed. This time was subtracted from the 'peak' time, to give the amount of time that was underestimated or overestimated, and for the purpose of this study is called the 'time to contact'.

Average time to contact (before and after peak time)
The condition which generated the least 'time to contact' (and was closest to the 'peak' time), was the *Image Only* condition (μ = 154.02 ms, SD = 681.05), followed by the *Audio-visual* condition (μ = 386.60 ms, SD = 548.87); and the *Audio Only* condition (μ = 443.59 ms, SD = 613.92).

Average time to contact (before peak time)
The condition that had the most number of trials in which the 'time to contact' was before the 'peak' time (therefore underestimating the contact time), was the *Audio Only* condition (with 25 trials, totaling 92.59% of the trials presented for that condition; weighted mean = 520.44 ms, weighted standard deviation = 391.87); and the *Audio-visual* condition, (with 24 trials, totaling 88.89% of the trials presented for that condition; weighted mean = 472.79 ms, weighted standard deviation = 249.79); followed by the *Image Only* condition (with 21 trials, totaling 77.78% of the trials presented for that condition; weighted mean = 324.94 ms, weighted standard deviation = 221.15).

Average time to contact (after peak time)
The condition that had the most number of trials in which the 'time to contact' was after the 'peak' time (therefore overestimating the contact time), was the *Image Only* condition (with 6 trials, totaling 22.22% of the trials presented for that condition; weighted mean = -170.93 ms, weighted standard deviation = 195.46); followed by the *Audio-visual* condition (with 3 trials, totaling 11.11% of the trials presented for that condition; weighted mean = -86.19 ms, weighted standard deviation = 94.92); and the *Audio Only* condition (with 2 trials, totaling 7.41% of the trials presented for that condition; weighted mean = -76.85 ms, weighted standard deviation = 35.30).

No trials had an average contact time during the image 'peak' (which had a duration of 41.67 ms, or one frame at 24 fps), with no individual participants indicating contact during this time.

4.5 Discussion

The results indicate that the *Image Only* condition had the slowest response to the contact time both before and after the peak time, with the least amount of underestimation before the 'peak' time and greatest amount of overestimation after the 'peak'.

However, the *Audio Only* condition, although still only providing unimodal information about the approaching object, prompted participants to have the fastest response to contact time overall, both before and after the peak image frame, with the greatest amount of underestimation before the peak time and least amount of overestimation after the peak. This suggests that the addition of sound and looming audio cues (in both the *Audio Only* condition and the *Audio-visual* condition) prompted people to underestimate the contact time more often, and with a greater time frame, than the scenes that had no sound.

5 Conclusion

Although the individual sound parameters that act as the audio cues for an approaching object could not be controlled and varied in this study, this investigation of the complex sounds in there original form as created by the sound designers has shown that the addition of sound, and the multiple techniques used to create audio cues, cause people to underestimate the contact time of an approaching object. This result suggests that further investigation is warranted, with future research on the complex stimulis' individual sound parameters, as independent variables, and the perception generated as a result.

References

1. Rosenblum, L, Carello, C, & Pastore, R. (1987). *Relative effectiveness of three stimulus variables for locating a moving sound source.* Perception, 16, 175186.
2. Neuhoff, J.G. (2001). *An adaptive bias in the perception of looming auditory motion.* Ecological Psychology, 13 (2), 87-110.
3. Neuhoff, J.G., & Heckel, T. (2004). *Sex differences in perceiving auditory looming produced by acoustic intensity change.* In Proceedings of ICAD 04-Tenth Meeting of the International Conference on Auditory Display, Sydney, Australia.
4. Cappe, C., Thut, G., Romei, V., & Murray, M. M. (2009). *Selective integration of auditory-visual looming cues by humans.* Neuropsychologia, 47, 1045-1052.
5. Wilkie, S., Stockman. T., & Reiss, J. D. (2012). *Amplitude Manipulation For Perceived Movement In Depth.* Audio Engineering Society, 132nd Convention, Budapest.
6. Ghazanfar, A.A., Neuhoff, J.G., & Logothetis, N.K. (2002). *Auditory looming perception in rhesus monkeys.* In Proceedings of the National Academy of Sciences, USA 99, 15755-15757.
7. Maier, J. X., Chandrasekaran, C., Ghazanfar, Asif A, Spemannstrasse, B. C., Germany, T., & Procedures, E. (2008). *Integration of Bimodal Looming Signals through Neuronal Coherence in the Temporal Lobe.* Current Biology, (18), 963-968.
8. Maier, J.X., & Ghazanfar. A.A. (2007). *Looming biases in monkey auditory cortex.* Journal of Neuroscience, 27 (15), 4093-4100.
9. Maier, J, Neuhoff, J, Logothetis, N, & Ghazanfar. A. (2004). *Multisensory integration of looming signals by rhesus monkeys.* Neuron, 43 (2), 177-181.
10. Neuhoff, John G. (2004). *Ecological psychoacoustics: introduction and history.* Ecological Psychoacoustics. Elsevier Academic Press, California.
11. Rosenblum, L, Wuestefeld, A., & Saldana, H. (1993). *Auditory looming perception: Influences on anticipatory judgements.* Perception, vol 22, 1467-1482.

Taking Advantage of Editorial Metadata to Recommend Music

Dmitry Bogdanov and Perfecto Herrera

Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
{dmitry.bogdanov,perfecto.herrera}@upf.edu

Abstract. In this work we propose a novel approach to music recommendation based exclusively on editorial metadata. To this end, we propose to use a public database of music releases *Discogs.com*, which contains extensive information about artists, their releases and record labels. We rely on an explicit set of music tracks provided by the user as evidence of his/her music preferences to construct a user profile suitable for distance-based music recommendation. We evaluate the proposed method against two purely metadata-based approaches and one approach partially based on audio content in a listening experiment with 27 participants. The results of subjective evaluation show that the proposed method is competitive to the state-of-the-art recommenders based on commercial metadata, while being easily implemented relying only on open public data.

Keywords: Music recommendation, user modeling, music similarity, editorial metadata, subjective evaluation

1 Introduction

The amount of music available digitally has overwhelmingly increased during the last decade following the growth of the Internet and music technology developments. Nowadays vast amounts of music are available for listeners' access, but still finding relevant and novel music is often a difficult task for them. Thereby, music listeners and music scholars strive for better recommendation systems to facilitate music search and retrieval.

In this context, music recommendation is a challenging topic in the Music Information Research (MIR) community. The state-of-the-art approaches to music recommendation are based on measuring music similarity between artists or particular tracks, and on user profiling, eliciting the information about music preferences. To this end, both metadata and audio content information can be used. Considering metadata, the state-of-the-art approaches to recommend music exploit user ratings, consumption and listening history, which are commonly used for collaborative filtering, and social tags extracted from social tagging services for music such as *Last.fm*[1] or mined from the web pages related to music content [1–6]. Current metadata-based approaches can perform satisfactorily

[1] http://last.fm

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

for listeners' needs when dealing with popular music. However, such approaches have disadvantages. Firstly, due the long-tail problem [2], a system may not have sufficient and correct metadata information for unpopular items. This can significantly limit the quality and the scope of recommendations or even make them completely impossible. Secondly, such approaches are cold-start prone and costly to maintain, requiring a large amount of user ratings, consumption or listening behavior to be processed for collaborative filtering, or large databases of tags. This information is expensive to obtain and maintain, and, moreover, is generally proprietary.

Alternatively, audio content information, extracted from the raw audio signal, can be applied for music recommendation. Such approaches are able to achieve performance close, or even comparable, to successful metadata-based approaches in terms of the relevance of recommendations [7–9], avoiding the problem of the long tail. Nevertheless, they are computationally costly and thus they require a large effort to build and maintain large-scale music collections.

Concerning user profiling, there exist approaches based on user models, which employ classification into interest categories using content-based information [10–13] or hybrid sources [14]. As well, distance-based[2] approaches, starting from a set of preferred items in a content-based vector space [15, 9], or more complex hybrid probabilistic approaches [16, 17] are proposed.

In the present work, we focus on distance-based music recommendation approaches. Moreover we consider a passive scenario, when recommendations are provided based on knowledge of user preferences rather than on manual user-specified query-by-example. We aim for a lightweight approach suitable for large-scale music collections, in particular containing the long-tail of artists and tracks, while working with publicly available data.

We propose a novel recommendation approach which is based exclusively on editorial metadata. To this end, we propose to use a public database of music releases, *Discogs.com*,[3] which contains extensive user-built information on artists, labels, and their recordings. We rely on an explicit set of music tracks provided by a user as evidence of his/her music preferences, the henceforth called *"preference set"*. We construct a user profile suitable for distance-based music recommendation using editorial metadata about the artists from the user's preference set. More concretely, for each artist we retrieve a descriptive tag cloud, containing information about particular genres, styles, record labels, years of release activity, and countries of release fabrication. We then employ latent semantic analysis [18] to compactly represent each artist as a vector, and match the user's preference set to a music collection to produce recommendations. We evaluate the proposed approach together with a number of baseline approaches in terms of subjective satisfaction ratings and calculated amount of novel relevant and known trusted recommendations on real listeners.

This paper is organized as follows: In Section 2 we describe the considered approaches. Firstly, the proposed approach working exclusively on editorial

[2] We pragmatically refer to any music similarity measure with the term "distance".
[3] http://discogs.com

metadata (Section 2.1). Secondly, a hybrid baseline approach, which employs content-based semantic distance followed by a simple genre refinement. Thirdly, a metadata-based baseline approach, working on artist tag annotations obtained from the *Last.fm*[4] social music service. Fourthly, a state-of-the-art commercial recommender on the example of *iTunes Genius*,[5] which relies on a collaborative filtering information. All three baseline approaches are described in Section 2.2. In Section 3 we present the subjective evaluation of the considered approaches conducted on 27 participants. Section 3.1 provides the characterization of subjects, while Section 3.2 explains the listening experiment instructions, stimuli and procedure. The evaluation results are presented and discussed in Section 3.3. Finally, we conclude this study in Section 4.

2 Studied Approaches

The approaches considered in this work are distance-based. We focus on the use-case of passive recommendations based on user preferences similarly to our previous works [9, 19]. Therefore, the approaches provide track recommendations from a given music collection (the henceforth called *music collection*) starting from a set of tracks, given by the user as evidence of her/her music preferences (a *preference set*), and applying distance measures between the tracks in preference set and the tracks in music collection. To create a preference set, the user is asked to provide a minimal set of music tracks, which she/he believes to be sufficient to grasp or convey her/his music preferences. The tracks can be submitted solely using the essential editorial metadata sufficient to identify them and, additionally, in audio format. The editorial metadata and audio for all provided tracks is then retrieved, if missing. Metadata is cleaned by means of tag cleaning and audio fingerprinting software MusicBrainz.[6] Thus, we obtain both editorial metadata and audio content for each track from the user's preference set which are suitable to apply both metadata-based and content-based analysis and recommendation procedures.

We employed a large in-house music collection as the source for recommendations. This collection covers a wide range of genres, styles, and arrangements, containing 68K music excerpts (30 sec.) by 16K artists with a maximum of 5 tracks per artist. For consistency, in our experiments we require each of the recommendation approaches to output 15 tracks by different artists (1 track per artist) not being present among the artists in the user's preferences set. To this end, each approach includes an artist filter.

[4] All tags were obtained on March, 2011.

[5] `http://www.apple.com/itunes/features/` all experiments were conducted using iTunes 10.3.1 on December, 2011.

[6] `http://musicbrainz.org/doc/MusicBrainz_Picard`

2.1 Proposed Approach: Artist Similarity Based On Editorial Metadata (M-DISCOGS)

The approach we proposed works exclusively on editorial metadata found in the *Discogs.com* database. The dump of this database is released under the Public Domain license[7], which makes is useful for different music applications, and in particular for research purposes of the MIR community. While there exist similar music services, such as public *MusicBrainz*[8] database, or proprietary *Last.fm* or AllMusic[9], we opt for *Discogs* as it contains the largest catalog of music releases and artists, while being known for accurate moderated metadata, which includes comprehensive annotations of particular releases.

The database contains the extensive information about up to 2,848K releases, 2,195K artists, and 281K labels.[10] In particular, for each artist this information includes a list of aliases, members (in the case an artist is a group), and group memberships (in the case an artist is a single person). Moreover it contains a list of releases authored by the artist, including albums, singles and EPs, and a list of appearances on the releases headed by other artists or compilations. A release corresponds to a particular edition of an album, single, EP, etc., and the releases related to the same album, single, or EP, can be grouped together into a "master release". Each release contains genre, style, country and year information. Genres are broad categories (such as classical, electronic, funk/soul, jazz, rock, etc.) while styles are more specific categories (such as neo-romantic, tech house, afrobeat, free jazz, viking metal, etc.) In total the database counts up to 15 genre categories and 329 styles.

For each artist in the database[11] we create a tag-cloud using genre, style, label, country, and year information related to this artist. To this end, we retrieve three lists of releases (*MAIN*, *TRACK*, *EXTRA*), where the artist occurs as (1) main artist, heading the release, (2) track artist, for example being on a compilation or with a guest appearance on a release, (3) extra artist, being mentioned in the credits of a release (usually related to the activity such as remixing, performing, writing and arrangement, production, etc.).

For each found release related to the artist, we retrieve genre, style, label, country, and year tags. For each of the three lists, we merge releases accordingly to their master release, keeping the genres, styles, and countries, which are present in at least one of the releases (i.e., applying a set union). Concerning the release years, we attempt to approximate the authentic epoch, when the music was firstly recorded, produced, and consumed. As a master release can contain reissues along with original releases, we keep the earliest (the original) year and, moreover, propagate it with descending weights as following:

$$W_{y\pm i} = W_y * 0.75^i, i \in \{1, 2, 3, 4, 5\} \quad (1)$$

[7] `http://www.discogs.com/data/`
[8] `http://musicbrainz.org`
[9] `http://www.allmusic.com`
[10] As on January 3, 2012.
[11] In our experiments, we used a *Discogs* monthly dump dated by January, 2011.

where W_y is the original year y, and 0.75 is a decay coefficient. For example, if the original year "y" is 1995, the resulting year-tag weights will be $W_{1995} = 1.0, W_{1994} = W_{1996} = 0.75, W_{1993} = W_{1997} \approx 0.56, W_{1992} = W_{1998} \approx 0.42, W_{1991} = W_{1999} \approx 0.32, W_{1990} = W_{2000} \approx 0.24$.

Thereafter, we summarize *MAIN*, *TRACK*, and *EXTRA* lists of the artist to a single tag-cloud. We assume a greater importance of tag annotations for the main artist role in comparison to track artists or extra artists; e.g., tags found on an artist's album are more important than the ones found on a compilation. We empirically assign the weights to these three groups of artist roles: 1.0 for main artists and 0.5 for both track and extra artists. As well, we assign further weights to tags according to their category: 1.0 for genres, styles, and labels, and 0.5 for years and countries, rescaling the artist tag-cloud. In particular, we decided to give equal importance to label information as to genres and styles. The rational behind grounds on the hypothesis that record label information gives a very valuable clue to a type of music, especially in the long-tail for the case of niche labels.

Finally, we propagate artist tags using the artist relations found in the database, such as aliases and membership relations. We suppose related artists to share similar musical properties and, therefore, assure that artists with low amount of releases will obtain reasonable amount of tags. To this end, for each artist we add a set of weighted tag-clouds of all related artists to the associated tag-cloud. We select a propagation weight of 0.5 and apply only 1-step propagation; i.e. tags will be propagated only between artists sharing a direct relation. Figure 1 presents an example of the proposed annotation procedure.

Following the described procedure we are able to construct tag-clouds for each artist in the *Discogs* database which together form a sparse tag matrix. To simplify this matrix, for each artist we apply additional filtering by means of erasing the tags with weight less than 1% of the artist's tag with the maximum weight. We then apply latent semantic analysis [18, 20, 21] to reduce the dimensionality of the obtained tag matrix to 300 latent dimensions. Afterwards, Pearson correlation distance [22, 2] can be applied on the resulting topic space to measure similarity between artists.

Once we have matched the annotated artists to the tracks in our music collection and the user's preference set, we retrieve recommendations applying the tag-based distance by the following procedure. For each track X in the user's preference set (a recommendation source), we apply this distance to retrieve the closest track C_X (a recommendation outcome candidate) from the music collection and form a triplet $(X, C_X, distance(X, C_X))$. We sort the triplets by the obtained distances, delete the duplicates of the recommendation sources (i.e., each track from the preference set produces only one recommendation outcome), and apply an artist filter. We return, as recommendations, the recommendation outcome candidates from the top 15 triplets. If it is impossible to produce 15 recommendations due to the small size of the preference set (less than 15 tracks) or because of the applied artist filter, we increase the number of possible recommendation outcome candidates per recommendation source.

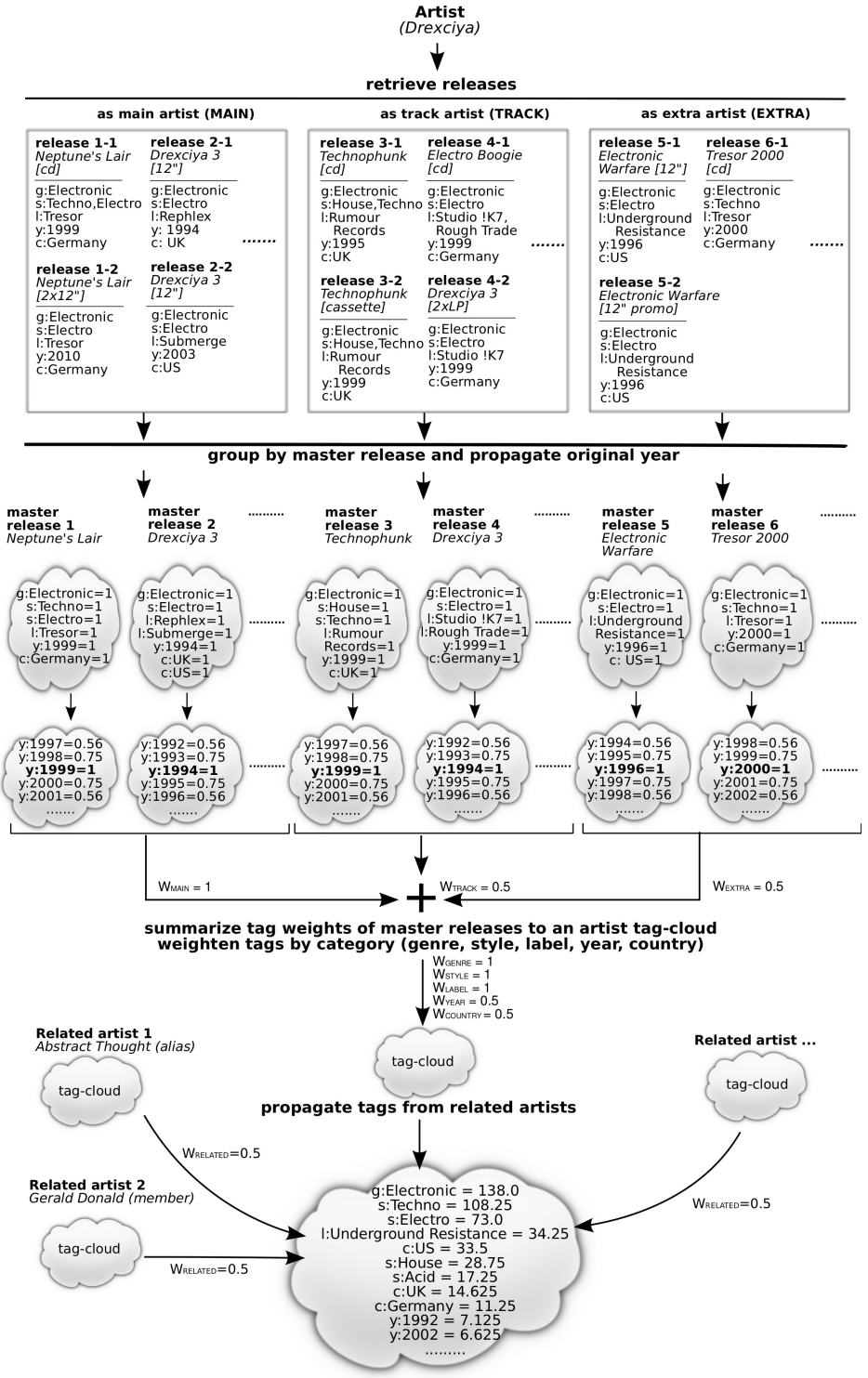


Fig. 1. An example of the proposed artist annotation based on editorial metadata from *Discogs*. Three lists of releases (MAIN, TRACK, EXTRA) are retrieved according to an artist's role. Particular releases are summarized into master releases, merging all found genre, style, label, and country tags, and selecting and propagating original year. Thereafter, tags are weighted to form a tag-cloud of an artist, and summed with the propagated tags of all related artists. Letters "g", "s", "l", "y", "c" stand for genre, style,

Pseudo-code of the distance-based recommendation procedure.

```
set IGNORE_ARTISTS to artists in preference set
remove tracks by IGNORE_ARTISTS from music collection
set N_OUTCOMES to 1
set N_RECS to 15

while true:
  set POSSIBLE_RECS to an empty list
  for track X in preference set:
    set X_NNS to N_OUTCOMES closest to X tracks in music collection
    for track C_X in X_NNS:
      append triple(X,C_X,distance(X,C_X)) to POSSIBLE_RECS
  sort POSSIBLE_RECS by increasing distance

  set RECS to an empty list
  for triple(SOURCE,OUTCOME,DISTANCE) in POSSIBLE_RECS:
    if OUTCOME occurs in RECS:
      next iteration
    if SOURCE occurs in RECS >= N_OUTCOMES times:
      next iteration
    append triple (SOURCE,OUTCOME,DISTANCE) to RECS
    if length of RECS list is N_RECS:
      return outcomes from RECS as recommendations
  set N_OUTCOMES to N_OUTCOMES + 1
```

2.2 Baseline Approaches

Content-based Semantic Similarity Refined By Genre Metadata (C-SEM+M-GENRE). As our first baseline, we consider a content-based semantic measure, providing a distance between tracks, filtered by genre metadata. The research presented in [19] has shown that simple filtering by a single genre tag can significantly improve the performance of a content-based-only approach to recommendation. Meanwhile, such genre information is considerably cheap to gather and maintain, it is however sufficiently descriptive for effective filtering.

We employ a semantic distance working on a set of high-level semantic descriptors (genres, musical culture, moods, instrumentation, rhythm, and tempo) inferred by support vector machines (SVMs) from low-level timbral, temporal, and tonal features. This distance has already been evaluated in the context of music similarity and music recommendation based on preference sets [23, 9, 24]. We refer the interested reader to the afore-cited literature for the implementation details of this measure and the evaluation results.

The semantic distance is applied similarly to the above-mentioned procedure for the M-DISCOGS, but in conjunction with a simple filtering: only the tracks of the same genre labels are considered as possible recommendation outcomes. We reproduce genre filtering as described in [19].

Artist Similarity Based On Last.fm Tags (M-TAGS) We consider a purely metadata-based similarity measure working on the artist level. This approach is based on social tags provided by the *Last.fm* API, retrieved for the artists from the user's preference set and the music collection. Using the API, we obtain a weight-normalized tag list for each artist. The weight ranges in the $[0, 100.0]$ interval, and we select a minimum weight threshold of 10.0 to filter out possibly inaccurate tags. The resulting tags are then assigned to each track in the preference set and the music collection. We then apply latent semantic analysis [18, 20] to reduce dimensionality to 300 latent dimensions. Pearson correlation distance [2] can be applied on the resulting topic space. We retrieve recommendations following the same procedure as for the M-DISCOGS.

Black-box Similarity By iTunes Genius (M-GENIUS) We consider commercial black-box recommendations obtained from the *iTunes Genius* playlist generation algorithm similarly to [8, 19]. Given a music collection and a query, this algorithm is capable to generate a playlist by means of an undisclosed underlying music similarity measure. It works on metadata and partially employs collaborative filtering of large amounts of user data (music sales, listening history, and track ratings) [8].

We randomly select 15 tracks, which are recognizable by *GENIUS*, from the user preference set. For each of the selected tracks (a recommendation source), we generate a playlist, apply the artist filter, and select the top track as the recommendation outcome. We increase the amount of possible outcomes per source when it is impossible to produce 15 recommendations similarly to the M-DISCOGS.

3 Evaluation

3.1 Subjects

We asked 27 voluntary subjects (selected from the authors' colleagues, their acquaintances and families) to provide their respective preference sets. Moreover, additional information was gathered, including personal data (gender, age, interest for music, musical background), and a description of the strategy and criteria followed to select the music pieces. The participants were not informed about any further usage of the gathered data, such as giving music recommendations.

The age of participants varied between 19 and 46 years ($\mu = 31.22$, $\sigma = 5.57$). All participants showed a very high interest in music (rating with $\mu = 9.43$ and $\sigma = 0.91$, where 0 means no interest and 10 means passionate). In addition, 24 participants play at least one musical instrument. The number of tracks selected by the participants to convey their musical preferences was very varied. It ranged from 8 to 178 music pieces ($\mu = 51.41$, $\sigma = 38.38$) with the median being 50 tracks. The time spent on creating a preference set differed a lot as well, ranging from 12 minutes to 60 hours ($\mu = 8.21$, $\sigma = 16.55$) with the median being two hours. The strategy followed by the participants to gather preference sets

also varied. Driving criteria for the selection of tracks included musical genre, mood, uses of music (listening, dancing, singing, playing), expressivity, musical qualities, and chronological order. We expect our population to represent a wide range of music enthusiasts, considering this information.

3.2 Evaluation Methodology

In general, evaluation of music recommender systems is a complicated, and, so far, not standardized procedure. Existing research works on music recommendation involving evaluations with real participants [10, 1, 25, 7, 8, 26] are significantly limited in the the tradeoff condition between the number of participants or by the number of evaluated tracks per approach by a particular user. Moreover, they are often focused on perceived quality of music similarity measures instead of user satisfaction with recommendations. In the latter case, evaluations generally include only one measure of user satisfaction, while the familiarity factor is rarely considered.

We describe our methodology of the conducted evaluation. Participants were asked to perform a blind subjective listening evaluation of the music generated using the 4 different recommendations approaches. To generate recommendations, we used our in-house music collection described in Section 2. For each participant, starting from her/his preference set, four recommendation playlists were generated by four respective approaches. Each playlist consisted of 15 tracks, and never contained more than one track by the same artist, nor contained tracks by artists from the preference set, due to the applied artist filter. All four playlists were then merged, randomized, and their filenames and metadata anonymized, and presented to a participant. This allowed to avoid any response bias due to presentation order, recommendation approach, or contextual recognition of tracks (e.g., by artist names). Furthermore, the participants were not aware of the amount of recommendation approaches, their names and rationales.

To gather feedback on recommendations, we provided a questionnaire for the subjects to express their subjective impressions related to the recommended music. To this end, we used four rating scales, following our previous works [9, 19]: A "*familiarity*" rating ranged from the identification of artist and title (4) to absolute unfamiliarity (0), with intermediate steps for knowing the title (3), the artist (2), or just feeling familiar with the music (1). A "*liking*" rating measured the enjoyment of the presented music with 0 and 1 covering negative liking, 2 being a kind of neutral position, and 3 and 4 representing increasing liking for the musical excerpt. A rating of "*listening intentions*" measured preference, but in a more direct and behavioral way than the "*liking*" scale, as an intention is closer to action than just the abstraction of liking. Again this scale contained 2 positive and 2 negative steps plus a neutral one. Finally, an even more direct rating was included with the name "*give-me-more*" allowing just 1 or 0 to respectively indicate a request for, or a reject of, more music like the one presented. We also asked users to provide title and artist for those tracks rated high in the familiarity scale. The textual meaning of the ratings was presented to the participants together with the allowed rating values.

3.3 Evaluation Results

First, we manually corrected the familiarity rating when the artist/title, provided by the participant, was incorrect. Hence, a familiarity rating of "3" or, more frequently, "4", was sometimes lowered to 1 or 2. These corrections represented 4.5% of the total familiarity judgments.

The four gathered ratings can be used to characterize different aspects of the considered recommendation approaches. We expect a good recommender system to provide high liking, listening intentions, and "give-me-more" ratings. Moreover, if we focus on music discovery, low familiarity ratings are desired, which will guarantee the novelty of relevant (liked) recommendations. Following [9, 19], we recoded the participants' ratings for each evaluated track into three categories which refer to the type of the recommendation: *hits*, *fails*, and *trusts*. We defined a recommended track to be a hit when it received low familiarity ratings (< 2) and high liking (> 2), listening intentions (> 2), and "give-me-more" ($= 1$) ratings simultaneously. Similarly, trusts were the tracks with high liking, listening intentions, "give-me-more", but as well high familiarity (> 1). Trusts, provided their overall amount is low, can be useful for a user to feel that the recommender is understanding his/her preferences [8] (i.e., a user could be satisfied by getting a trust track from time to time, but annoyed if every other track is a trust). Fails were the tracks which received low liking (< 3), listening intentions (< 3) and "give-me-more" ($= 0$) ratings. In any other case (e.g., a track received high liking, but low listening intentions and "give-me-more") the outcome category was considered to be "unclear", amounting to 17.3% of all recommendations.

We report the percent of hit, fail, trust, and unclear outcomes per recommendation approach in Table 1. According to the results of a chi-square test, an association between the approaches and the outcome categories ($\chi^2(9) = 46.879$, $p < 0.001$) can be accepted. Namely, certain approaches provide hits, fails or trust percents which are statistically different than what a flat distribution (i.e., equiprobable) would yield.

In general, the proposed M-DISCOGS approach performed well comparing to the baselines. The M-DISCOGS provided a considerably low (34.4%) amount of fails, being in between of the metadata-based baselines M-TAGS (with the lowest amount of fails, 32.8%) and M-GENIUS. In contrast, the C-SEM+M-GENRE approach, which is partially content-based, provided the largest (over 41%) amount of fails. Considering hits, the M-TAGS (38.8%) and C-SEM+M-GENRE (37.9%) are the leaders followed by M-GENIUS, and lastly, the M-DISCOGS. That is, our proposed approach provided the least amount of novel relevant recommendations (31.9%). Nevertheless this fact is compensated by the largest amount of trusts, gathered by the M-DISCOGS (16.4%) followed by the M-GENIUS (13.2%), M-TAGS, and the C-SEM+M-GENRE (4.4%). The amount of unclear recommendations ranged as well. As such recommendations consisted of the tracks with inconsistent ratings, we may not expect such tracks to be as relevant as hits and trust categories. Still, such tracks can be useful for certain scenarios (e.g., playlist generation), but are probably not well suited for others (e.g., digital music vending). Considering the extreme case, when

Table 1. Percent of fail, trust, hit, and unclear categories per recommendation approach.

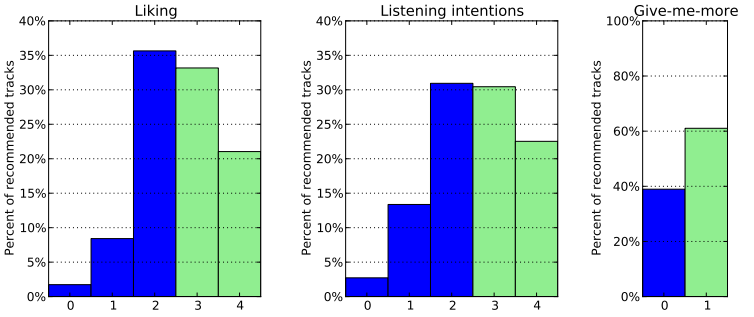
Approach	fail	hit	trust	unclear	hit+trusts
M-TAGS	32.8	38.8	7.4	21.0	46.2
M-DISCOGS	34.4	31.9	16.4	17.3	48.3
M-GENIUS	36.2	35.7	13.2	14.9	48.9
C-SEM+M-GENRE	41.6	37.9	4.4	16.1	42.3

Table 2. Mean ratings per recommendation approach.

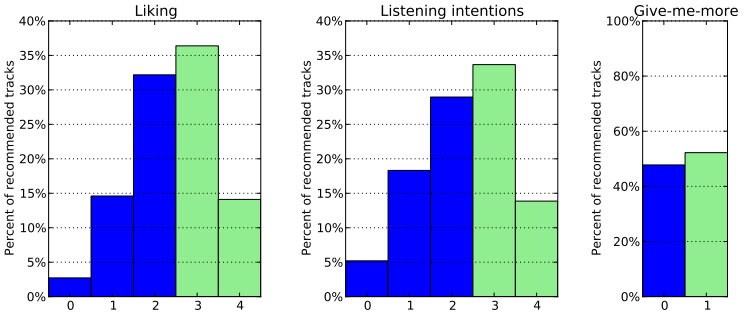
Approach	liking	listening intentions	give-me-more	familiarity
M-DISCOGS	2.63	2.57	0.63	0.83
M-GENIUS	2.60	2.50	0.59	0.80
M-TAGS	2.52	2.45	0.63	0.49
C-SEM+M-GENRE	2.45	2.33	0.52	0.37

fails and unclear categories are both unwanted outcomes, the metadata-based M-GENIUS and M-DISCOGS result as approaches with the least amount of unwanted recommendations (51.1% and 51.7%, respectively), followed by the M-TAGS, and lastly by the partially content-based C-SEML+M-GENRE approach (57.7%). In contrast, considering trusts and hits as wanted outcomes, the M-GENIUS and M-DISCOGS provide their largest amount (48.9% and 48.3%, respectively), followed by the M-TAGS and C-SEM+M-GENRE.

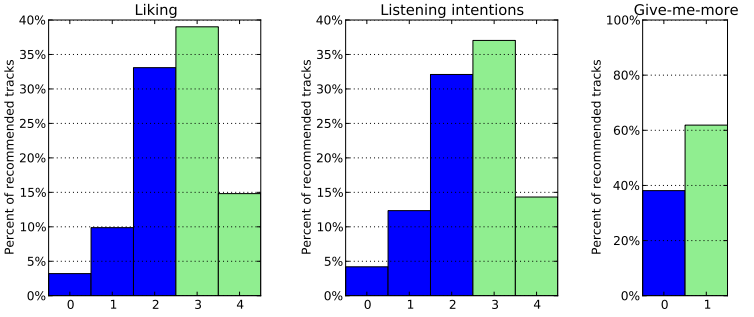
Apart from analysis of the outcome categories, we tested the effect of the recommendation approaches on the liking, listening intentions, and "give-me-more" subjective ratings. To this end, we conducted three separate between-subjects ANOVAs. Tested approaches were shown to have an impact on these ratings ($F(3, 1612) = 3.004$, $p < 0.03$ for the liking rating, $F(3, 1612) = 3.660$, $p < 0.02$ for the intentions rating, and $F(3, 1612) = 3.363$, $p < 0.02$ for the "give-me-more" rating). Pairwise comparisons using Tukey's test revealed differences only between M-DISCOGS vs C-SEM+M-GENRE for the case of all three ratings, and, in addition, a difference between M-TAGS vs C-SEM+M-GENRE in the case of the "give-me-more" rating. In Figure 2 we present the histograms for the liking, listening intentions, and "give-me-more" ratings. Mean values of these ratings are provided in Table 2. Inspecting the means, we see that all considered approaches performed with a user satisfaction slightly above average. Almost half of the provided recommendations were favorably evaluated, i.e., received high liking and listening intentions ratings (> 2) and a positive "give-me-more" request. An inspection of histograms shows that the proposed M-DISCOGS approach receives the highest amount of maximum ratings for liking and listening intentions ($\simeq 21\%$ and $\simeq 22.5\%$, respectively). In contrast, the amount of received negative ratings is lower. Still, returning to the ANOVA results, the only clear difference in performance, as measured by our 3 indexes, happens between M-



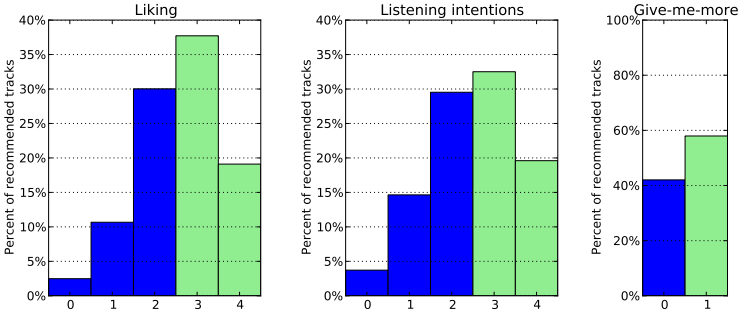
(a) M-DISCOGS



(b) C-SEM+M-GENRE



(c) M-TAGS



(d) M-GENIUS

Fig. 2 Histograms of liking, listening intentions, and "give me more" ratings gathered

DISCOGS and C-SEM+M-GENRE. In other words, the proposed M-DISCOGS approach is able to achieve similar liking, listening intentions and willingness to get recommended music as existing (and commercial) state-of-the-art systems. Interestingly, we have evidenced the above-average ceiling in the performance of all considered approaches. This fact highlights a lot of room for improvement of music recommender systems.

4 Conclusions

We have considered and evaluated different distance-based approaches to music recommendation, starting from a set of music tracks explicitly provided by a user as an evidence of his/her musical preferences. We proposed a novel approach working exclusively on editorial metadata taken from publicly available music database, *Discogs.com*. Relying on user-built information about music releases present in this database, we demonstrated how this information can be applied to create descriptive tag-based artist profiles, containing information about particular genres, styles, record labels, years of release activity, and countries. Furthermore, to overcome the problem of tag sparsity, such artist profiles can be compactly represented as vectors in a latent semantic space of reduced dimension. Applying a distance measure between the resulting artist vectors for the tracks in the preference set of a user and the tracks within a music collection, we are able to generate recommendations.

The proposed approach has a number of advantages over common metadata-based approaches. Firstly, our approach is able to provide a compact profile for each artist found in *Discogs* database. Matching these profiles to music collections, large-scale recommendation systems can be built. Secondly, the proposed approach is based only on open public data, meanwhile the majority of successful recommender systems operate on commercially withhold metadata. As a consequence, our approach is easy to create and reproduce. Subjective evaluation of the proposed approach with 27 participants demonstrated performance comparable to the state-of-the-art metadata-based approaches, including an industrial recommender. In particular, our approach provided large amount of trusted and novel relevant recommendations, which suggests that the proposed approach is well suited for music discovery and playlist generation. Although we have considered and evaluated the proposed approach in the context of "passive discovery", relying on preference sets provided by listeners, we expect our conclusions to be applicable for the query-by-example use-case.

Interestingly, the evaluated content-based approach filtered by simple genre metadata revealed performance comparable to the metadata-based approaches as well. In our previous research [9, 19], we evidenced a high number of trusted recommendations for the metadata-based approaches, and fewer in the case of content-based recommendations similarly to the present study. Moreover, we similarly evidenced the user satisfaction by the evaluated metadata-based approaches to be slightly above average showing a lot of room for improvement.

Future work will be focused on the limitations of the current proof-of-concept study. A number of parameters were chosen empirically in the proposed approach and will require further research to find optimal weights for the release types, tag types, and artist propagation as well as the year propagation decay. Moreover, a hybrid approach expanding the proposed method with audio content information will be of interest.

Acknowledgments. The authors would like to thank all participants involved in the evaluation. This research has been partially supported by the FI Grant of Generalitat de Catalunya (AGAUR) and the Classical Planet (TSI-070100-2009-407, MITYC), DRIMS (TIN2009-14247-C02-01, MICINN), and MIRES (EC-FP7 ICT-2011.1.5 Networked Media and Search Systems, grant agreement No. 287711) projects.

References

1. Firan, C.S., Nejdl, W., Paiu, R.: The benefit of using tag-based profiles. In: Latin American Web Conf. (2007) 32–41
2. Celma, O.: Music recommendation and discovery in the long tail. PhD thesis, UPF, Barcelona, Spain (2008)
3. Baltrunas, L., Amatriain, X.: Towards time-dependant recommendation based on implicit feedback. In: Workshop on Context-aware Recommender Systems (CARS'09). (2009)
4. Jawaheer, G., Szomszor, M., Kostkova, P.: Comparison of implicit and explicit feedback from an online music recommendation service. In: Int. Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec'10). HetRec '10, New York, NY, USA, ACM (2010) 47–51 ACM ID: 1869453.
5. Levy, M., Bosteels, K.: Music recommendation and the long tail. In: ACM Conf. on Recommender Systems. Workshop on Music Recommendation and Discovery (Womrad 2010). (2010)
6. Schedl, M., Pohle, T., Knees, P., Widmer, G.: Exploring the music similarity space on the web. ACM Trans. on Information Systems **29** (2011) 1–24
7. Magno, T., Sable, C.: A comparison of signal-based music recommendation to genre labels, collaborative filtering, musicological analysis, human recommendation, and random baseline. In: Int. Conf. on Music Information Retrieval (ISMIR'08). (2008) 161–166
8. Barrington, L., Oda, R., Lanckriet, G.: Smarter than genius? human evaluation of music recommender systems. In: Int. Society for Music Information Retrieval Conf. (ISMIR'09). (2009) 357–362
9. Bogdanov, D., Haro, M., Fuhrmann, F., Gómez, E., Herrera, P.: Content-based music recommendation based on user preference examples. In: ACM Conf. on Recommender Systems. Workshop on Music Recommendation and Discovery (Womrad 2010). (2010)
10. Hoashi, K., Matsumoto, K., Inoue, N.: Personalization of user profiles for content-based music retrieval based on relevance feedback. In: ACM Int. Conf. on Multimedia (MULTIMEDIA'03). (2003) 110–119

11. Grimaldi, M., Cunningham, P.: Experimenting with music taste prediction by user profiling. In: ACM SIGMM Int. Workshop on Multimedia Information Retrieval (MIR'04). (2004) 173–180
12. Moh, Y., Orbanz, P., Buhmann, J.M.: Music preference learning with partial information. In: IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. (2008) 2021–2024
13. Moh, Y., Buhmann, J.M.: Kernel expansion for online preference tracking. Proceedings of The Int. Society for Music Information Retrieval (ISMIR) (2008) 167–172
14. Su, J.H., Yeh, H.H., Tseng, V.S.: A novel music recommender by discovering preferable perceptual-patterns from music pieces. In: ACM Symp. on Applied Computing (SAC'10). (2010) 1924–1928
15. Logan, B.: Music recommendation from song sets. In: Int. Conf. on Music Information Retrieval (ISMIR'04). (2004) 425–428
16. Yoshii, K., Goto, M., Komatani, K., Ogata, T., Okuno, H.G.: Hybrid collaborative and content-based music recommendation using probabilistic model with latent user preferences. In: Int. Conf. on Music Information Retrieval (ISMIR'06). (2006)
17. Li, Q., Myaeng, S.H., Kim, B.M.: A probabilistic music recommender considering user opinions and audio features. Information Processing & Management **43** (2007) 473–487
18. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American society for information science **41** (1990) 391–407
19. Bogdanov, D., Herrera, P.: How much metadata do we need in music recommendation? a subjective evaluation using preference sets. In: Int. Society for Music Information Retrieval Conf. (ISMIR'11). (2011) 97–102
20. Levy, M., Sandler, M.: Learning latent semantic models for music from social tags. Journal of New Music Research **37** (2008) 137–150
21. Sordo, M., Celma, O., Blech, M., Guaus, E.: The quest for musical genres: Do the experts and the wisdom of crowds agree? In: Int. Conf. of Music Information Retrieval (ISMIR'08). (2008) 255–260
22. Gibbons, J.D., Chakraborti, S.: Nonparametric Statistical Inference. CRC Press (2003)
23. Bogdanov, D., Serrà, J., Wack, N., Herrera, P.: From low-level to high-level: Comparative study of music similarity measures. In: IEEE Int. Symp. on Multimedia (ISM'09). Int. Workshop on Advances in Music Information Research (AdMIRe'09). (2009) 453–458
24. Bogdanov, D., Serrà, J., Wack, N., Herrera, P., Serra, X.: Unifying low-level and high-level music similarity measures. IEEE Trans. on Multimedia **13** (2011) 687–701
25. Kim, J., Jung, K., Ryu, J., Kang, U., Lee, J.: Design of ubiquitous music recommendation system using MHMM. Volume 2. (2008) 369–374
26. Lu, C., Tseng, V.S.: A novel method for personalized music recommendation. Expert Systems with Applications **36** (2009) 10035–10044

Oral session 7:
Computational Musicology and Music Education

Bayesian MAP estimation of piecewise arcs in tempo time-series

Dan Stowell[1] and Elaine Chew[1]

Centre for Digital Music, Queen Mary, University of London
dan.stowell@eecs.qmul.ac.uk

Abstract. In musical performances with expressive tempo modulation, the tempo variation can be modelled as a sequence of tempo arcs. Previous authors have used this idea to estimate series of piecewise arc segments from data. In this paper we describe a probabilistic model for a time-series process of this nature, and use this to perform inference of single- and multi-level arc processes from data. We describe an efficient Viterbi-like process for MAP inference of arcs. Our approach is score-agnostic, and together with efficient inference allows for online analysis of performances including improvisations, and can predict immediate future tempo trajectories.

Keywords: tempo, expression, Viterbi, time series

1 Introduction

In various types of musical performance, one component of the musical expression is conveyed in the short-term manipulation of tempo, with tempo modulation reflecting musical phrase structure [7, 9]. This has motivated various authors to construct automatic analyses of the arc-shaped tempo modulations in recorded musical performances, with or without score-derived information to supplement the analysis [7, 9, 5]. (See also [6] who fit piecewise linear arcs to rock and jazz data, applying similar techniques but to genres in which the underlying tempo is held more fixed.)

Machine understanding of tempo, including its variability, can be useful in live human-machine interaction [1, 8]. However most current online tempo-tracking systems converge to an estimate of the current tempo, modelling expressive variations as deviations rather than as components of an unfolding tempo expression. In this paper we work towards the understanding of tempo arcs in a real-time system, paving the way for automatic accompaniment systems which follow the expressive tempo modulation of players in a more natural way.

We also consider tempo arcs within a probabilistic framework. Previous authors have approached piecewise arc estimation using Dynamic Programming (DP) with cost functions based on squared error [5, 6]. These are useful and can provide efficient estimation, but by setting the problem in a probabilistic framework (and providing the corresponding Viterbi-like DP estimator), we gain some

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

advantages: prior beliefs about the length and shape of arcs can be expressed coherently as prior distributions; measurement noise is explicitly modelled; and the goodness-of-fit of models is represented meaningfully as posterior probabilities, which allows for model comparison as well as integration with other workflow components which can make use of estimates annotated with probability values. Note that while we describe a fully probabilistic model, for efficient inference we will develop a Maximum A Posteriori (MAP) estimator, which returns only the maximum probability parameter settings given the priors and the data.

In the following we will describe our model of arcs in time-series data, and develop an efficient MAP estimation technique based on least-squares optimisation and Viterbi-like DP. The approach requires some kind of unsmoothed instantaneous tempo estimate as its input, which may come from a tempo tracker or from a simple measurement such as inter-onset interval (IOI). We will then discuss how the estimator can be used for immediate-future tempo prediction, and how it can be applied to multiple levels simultaneously. Finally we will apply the technique to tempo data from three professional piano performances, and discuss what the analysis reflects in the performances.

2 Modelling and Estimation

For our basic model, we consider tempo to evolve as a function of metrical position (beat number) x in a musical piece as a series of connected arcs, where each arc's duration, curvature and slope are independently drawn from prior distributions (to be described shortly). Our model is deliberately simple, and agnostic of any score information that might be available. To sample from this model, we pick an initial tempo at the starting time, then define a single upwards tempo arc which starts from that point, and the tempo trajectory (speeding up and then slowing down) over a number of measures. Any tempo data which may be measured during this interval is modelled as being drawn from the arc plus some amount of gaussian noise. Once the ending breakpoint of this arc is reached, the next arc is sampled from the same priors, using the ending tempo as the new starting tempo. Hence each tempo arc is conditionally independent of all previous observations once the starting tempo is determined, i.e. once the previous arc's parameters are fixed. This assumption of conditional independence is slightly unrealistic, since it ignores long-range relationships between tempo arcs, but it accounts for the most important interactions and makes inference tractable.

Our basic model is also only single-level, assuming that a single arc contributes to the current tempo at any moment, rather than considering for example contributions from multiple timescales such as piece-level, movement-level, phrase-level and bar-level combined. In Section 2.4 we will consider a simple multi-scale extension of our technique, which we will apply in our analysis of piano performance data. (For an alternative approach in which various components can be simultaneously active see [7].)

2.1 Fitting a Single Arc

To fit a single arc shape to data, one can use standard quadratic regression, fitting a function of the form

$$f(x) = a + bx + cx^2, \tag{1}$$

and minimising the L_2 prediction error over the supplied data for $y \approx f(x)$. In the Bayesian context, we wish to incorporate our prior beliefs about the regression parameters (here a, b and c), which is related to the optimisation concept of *regularisation*, the class of techniques which aims to prevent overfitting by favouring certain parameter settings. In fact, a gaussian prior on a regression parameter can be shown to be equivalent to the conventional L_2-norm regularisation of the parameters [2, p. 153], summarised as:

$$\text{regularisation coefficient} = \frac{\text{variance of gaussian noise}}{\text{variance of gaussian prior}}. \tag{2}$$

This equivalence is useful because it allows us to use common convex optimisation algorithms to perform the equivalent regularised least squares optimisation, and they will yield the MAP estimate for the probabilistic model.

However, in this context a standard gaussian prior is not exactly what we require, since we are expecting upwards arcs and not troughs – we are expecting c in Equation 1 to be negative. A more appropriate choice of prior might be a negative log-gaussian distribution, which allows us to specify a "centre of mass" for the arc shapes (expressed through the log-mean and log-standard-deviation parameters), yet better represents our expectation that tempo arcs will always have negative curvature, (almost) flat and extremely strongly curved arcs being equally rare.

The unconventional choice of prior might seem to remove the equivalence of the MAP regression technique with standard regularised least squres. Yet if we rewrite our function to be

$$f(x) = a + bx - e^c x^2, \tag{3}$$

then our prior belief about this modified parameter c becomes a gaussian, yielding a negative-log-gaussian in combination with our function. In addition, we will use a standard gaussian prior on b. We could do the same for a but instead we will use an improper uniform prior, for reasons which will be described in Section 2.2. Therefore, our priors for Equation 3 will be gaussian priors on b and c, which can easily be converted to the equivalent L_2-regularisation terms for optimisation.

The strength of the regularisation (the value of the regularisation coefficient) reflects the specificity of our priors versus our data – specifically, the regularisation parameter is given by the noise variance divided by the prior variance [2, p. 153]. Again, we see how the probabilistic setting helps to ground our problem, connecting the strength of the regularisation directly to our prior beliefs about the model and the data rather than manually-tuned parameters.

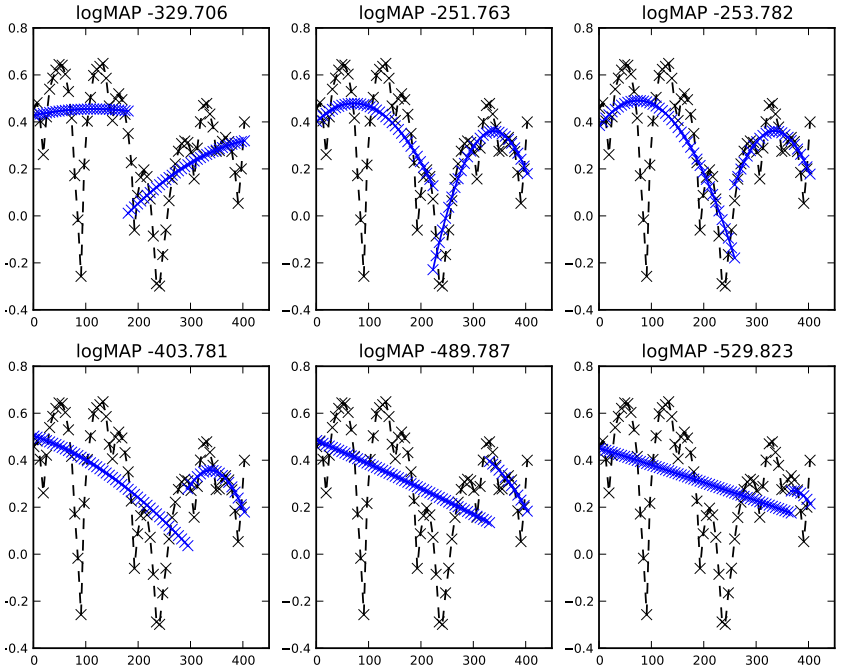


Fig. 1. A selection of piecewise arc fits performed on a synthetic dataset, with manually-specified breakpoint locations. The "logMAP" (log of MAP probability) values quoted with each one indicate the relative likelihood assigned to each fit, given the prior parameters chosen. (Prior parameters are the same for each of these plots.) The best-fitting plots have correspondingly higher (less negative) logMAP values.

2.2 Fitting Multiple Arcs

If a time-series is composed of multiple arcs and the breakpoints are known, then fitting multiple arcs is as simple as performing the above single-arc fit for each subsection of the time series (as in Figure 1). Additionally, one should take care of the arc's dependence upon its predecessor (to enforce that they meet up), which is not shown in these plots. In our case, we want to estimate the breakpoint locations as well as the arc shapes between those breakpoints. This can be performed by iterating over all possible combinations of one breakpoint, two breakpoints, three (...) for the dataset, and choosing the result with the lowest cost (the highest posterior likelihood).

The Bayesian setting makes it possible to compare these different alternatives (e.g. one single arc vs. one arc for every datapoint) without having to add arbitrary terms to counter overfitting; instead, we specify a prior distribution over the arc durations, which in combination with the other priors and data likelihoods yields a MAP probability for any proposed set of arcs. In this paper we choose a log-normal prior distribution over arc durations. See Figure 1 for

some examples of different sets of arcs fitting to a synthetic dataset, and the posterior (log-)probabilities associated.

In order for only a single tempo value to exist at each breakpoint (and not a discontinuous leap from one tempo to another), we fit each arc under the constraint that its starting value equals the ending value of the previous arc. This removes one degree of freedom from the function to be fit (Equation 3) which otherwise has three free parameters. We implement this by constraining the value of a in the optimisation so that the function evaluates to the predetermined value at the appropriate time-point. The least-squares optimisation therefore only operates on b and c.

2.3 Viterbi-like Algorithm

The number of possible combinations of arcs for even a small time-series (such as Figure 1) grows quickly very large, and so it is impractical to iterate all combinations. This is where Dynamic Programming (DP) can help. Here we describe our DP algorithm, which, like the well-known Viterbi algorithm, maintains a record of the most likely route that leads to each of a set of possible states. Rather than applying it to the states of a Hidden Markov Model, we apply it to the possibility that each incoming datum represents a breakpoint.

Assume that the first incoming datum is a breakpoint. (This assumption can be relaxed, in a similar way to the treatment of the final datum which we consider later.) Then, for each incoming datum $(x_n,\, y_n)$, we find what would be the most likely path *if it were certainly* a breakpoint. We do this by finding the most appropriate past datum $(x_{n-k},\, y_{n-k})$ which could begin an arc to the current datum – where the appropriateness is judged from the MAP probability of said arc, combined with the MAP probability of the whole multiple-arc history that leads up to that past datum (recursively defined).

With our lognormal prior on the arc lengths (and with many common choices of prior), the probability mass is concentrated at an expected time-scale, and very long arcs are highly improbable *a priori*. Hence in practice we truncate the search over potential previous arc points to some maximum limit K (i.e. $k \leq K$).

Thus, for every incoming data point we perform no more than K single-arc fits, then store the details of the chosen arc, the MAP probability so far, and a pointer back to the datapoint at the start of the chosen single arc. The simplest way to choose the overall MAP estimate is then to pick another definite breakpoint (for example, the last datum if the performance has finished) and backtrack from there to recover the MAP arc path.

Complexity The time complexity of the algorithm depends strongly on that of the convex optimisation used to perform a single-arc fit. Assume that the complexity of a single-arc fit is proportional to the number of data points k included in the fit, where $k \leq K$. Then for each incoming data point a search is performed for one subset each of 2, 3, ... K data points, which essentially yields an order $\mathcal{O}(K^2)$ process. For online processing this is manageable if K is

not too large. Analysing a whole dataset of M points then has time complexity $\mathcal{O}(K^2M)$. (Compare this to the broadly similar complexity analysis of [6].) The space complexity is simply $\mathcal{O}(M)$, or $\mathcal{O}(K)$ if the full arc history since the very beginning does not need to be stored. This is because a small fixed amount of data is stored per datapoint.

Predicting Immediate Future Arcs As discussed, if we know the performance has finished then we can find the Viterbi path leading to a breakpoint at the final data point received. However, we would also like to determine the most likely set of arcs in cases where the performance might not have finished (e.g. for real-time interactive systems), and thus where we do not wish to assert that the latest datum is a breakpoint. We wish to be able to estimate an arc which may still be in progress. If we can, this has a specific benefit of predicting the immediate future evolution of the tempo modulations (until the end of the present arc), which may be particularly useful for real-time interaction.

We can carry this out in our current approach as follows. Since an arc's duration (as well as the curve-fit) affects its MAP probability, in the case where the latest arc may or may not be terminating we must iterate over the arc's possible durations and pick the most likely. To do this we choose a set of future time-points as candidate breakpoints, $x_{n+1} \ldots x_{n+J}$ (e.g. an evenly-spaced tatum grid of $J = K$ future points). Then we supply these data to the Viterbi update process exactly as is done with actual data, but with no associated y values. These "hypothetical" Viterbi updates will use these time-points to determine the arc-lengths being estimated, and in normalising the data subset, but will not include them in the arc-fitting process. It will therefore yield a MAP probability estimate for each of the time-points as if an arc extended from the real data as far as this hypothetical breakpoint. Out of these possibilities, the one with the highest MAP probability is the MAP estimate for an arc which includes the latest real datum and some portion of the hypothetical future points. (The hypothetical Viterbi updates are not preserved: if more data comes in, it is appended to the Viterbi storage corresponding only to the actual data.)

2.4 Multi-scale Estimation

The model we describe operates at one level, with expected arc durations given by the corresponding prior. Our model is adaptable to any time-scale by simply adapting the prior. It does not however automatically lend itself to simultaneous consideration of multiple active timescales.

Multi-scale analysis can be carried out by analysing a dataset with one timescale, then analysing the residual at a second timescale. This residual-based decomposition has been used previously in the literature (e.g. [9]); it requires a strong hierarchical assumption that the arcs at the first timescale do not depend at all on those at the second timescale, while the second is subordinate to the first. We consider this to be unrealistic, since there may well be interactions between the different timescales on which a performer's expression evolves. However this assumption leads to a tractable analysis.

Note also that this approach to multi-scale estimation requires the first analysis to be completed (so that the residual is known) before the second scale can be analysed. Some DP approach may be possible to enable both to be calculated online, but we have not developed that here. For the present work, the single-scale Viterbi tracking is applicable and useful for online tracking, while multi-scale analysis is an offline process, which we will next apply to modelling of pre-recorded tempo data.

3 Analysis of Expressive Piano Performance

We applied our analysis to an existing set of annotations of three performances of Beethoven's *Moonlight Sonata*. The annotations were provided by Elaine Chew and have previously been analysed by Chew with reference to observations noted by Jeanne Bamberger [4]. For each of three well-known performances of the piece (by Daniel Barenboim (1987), Maurizio Pollini (1992) and Artur Schnabel (2009)), the first 15 bars have been annotated with note onset times, which correspond to regular triplet eighth-note timings.

We implemented the algorithm in Python, using the `scipy.optimize.fmin` optimiser to solve individual regressions. Source code is available.[1] (Note that this development implementation is not generally fast enough for real-time use.)

Instantaneous tempo was derived from these inter-onset intervals, then analysed using a two-pass version of our algorithm: first the data was analysed using an arc-duration prior centred on four bars; then the residual was analysed using an arc-duration prior centred on one bar. This choice of timescales is a relatively generic choice which might reasonably be considered to reflect a performer's short-term and medium-term state; however it might also be said to be a form of basic contextual information about the relevant timescales in the current piece. For the current study, we confine ourselves to priors with log-normal shapes, though an explicitly score-derived or corpus-derived prior could have a more tailored and perhaps multimodal shape.

Figure 2 shows the results. The analyses show some notable similarities and differences, some of which we will now discuss.

The longer time scale analysis (centred on four-bar durations) immediately highlights a difference between the performances: Pollini's performance appears to contain relatively little variation on this level, as the fit yields long and shallow arcs, with breakpoints near positions 48, 96 and 168 (structurally important positions; 96 is where the key-change occurs). On the other hand, both Barenboim and Schnabel's tempo curves exhibit fairly deep and varied arcs. Schnabel's performance exhibits the most dramatic variation in the first four bars until around measure 48: this first four-bar section corresponds to the opening statement of the basic progression, before the melody enters in the fifth bar (and the underlying progression repeats). Bamberger described Schnabel as performing them "as if in one long breath" (quoted in [4]), not quite reflected in our analysis.

[1] `https://code.soundsoftware.ac.uk/projects/arcsml`

On the shorter time scale, the analysis tends to group phrases into one-bar or two-bar arcs. Aspects of the musical structure are reflected in the arcs observed. Sections of the melody which lend themselves to two-bar phrasing (e.g. 72–96) are generally reflected in longer arcs crossing bar lines. Conversely, in the region 96–132 the change to the new key unfolds as each new chord enters at the start of a bar, and the tempo curves for all three performers reflect an expressive focus on this feature, with one-bar arcs which are more closely locked to the bar-lines than elsewhere. Note that in this section Schnabel matches Pollini in exhibiting a long and shallow arc on the slow timescale, with all the expressive variation concentrated on the one-bar arcs.

Over the excerpt generally, the breakpoints for Schnabel are further away from the barline than the others, as was observed in Chew's manual analysis.

We have extended the plots slightly beyond the 180 annotated data points, to illustrate the immediate-future predictions made by the model. (This is done for both timescales, though only the longer timescale (in red) shows noticeable extended arcs.) All the performers, and especially Schnabel, exhibit an acceleration towards the end of the annotated data, reflected in the predictions of an upward arc followed by a gradual slowing over the next bar. This type of prediction is plausible for such expressively-timed music.

To illustrate the effect that the prior parameters have upon the regression, Figure 3 shows the same analysis as Figure 2 but with the standard deviation of the noise prior set at 4.0 rather than 3.0. The increase in the assumed noise variance leads the algorithm to "trust" the data less and the prior slightly more (cf. Equation 2). In our example, some of the breakpoints for the long-term arcs (in red) have changed, losing some detail, though most of the detail of the second-level analysis (in blue) is consistent.

4 Conclusions

We have described a model with similarities to some previous piecewise-arc models of musical expression, but with a Bayesian formulation which facilitates model comparison and the principled incorporation of prior beliefs. We have also described an efficient Viterbi-like Dynamic Programming approach to estimation of the model from data. The approach provides scope to apply the model to real-time score-free performance tracking, including prediction of immediate future tempo modulation. Source code for the algorithm (in Python) is available.

We have applied the model in a two-level analysis to data from expressive piano performance, illustrating the algorithm's capacity to operate at different time-scales, and to recover expressive arc information that corresponds with some musicological observations regarding phrasing and timing.

Further research would be needed to develop a model considering multiple simultaneously-active levels of expression which can be applied online as with our single-level Viterbi-like algorithm. Similar arcs have been observed and analysed in loudness information extracted from performances [3]. It would also be useful to combine loudness information with tempo information in this model.

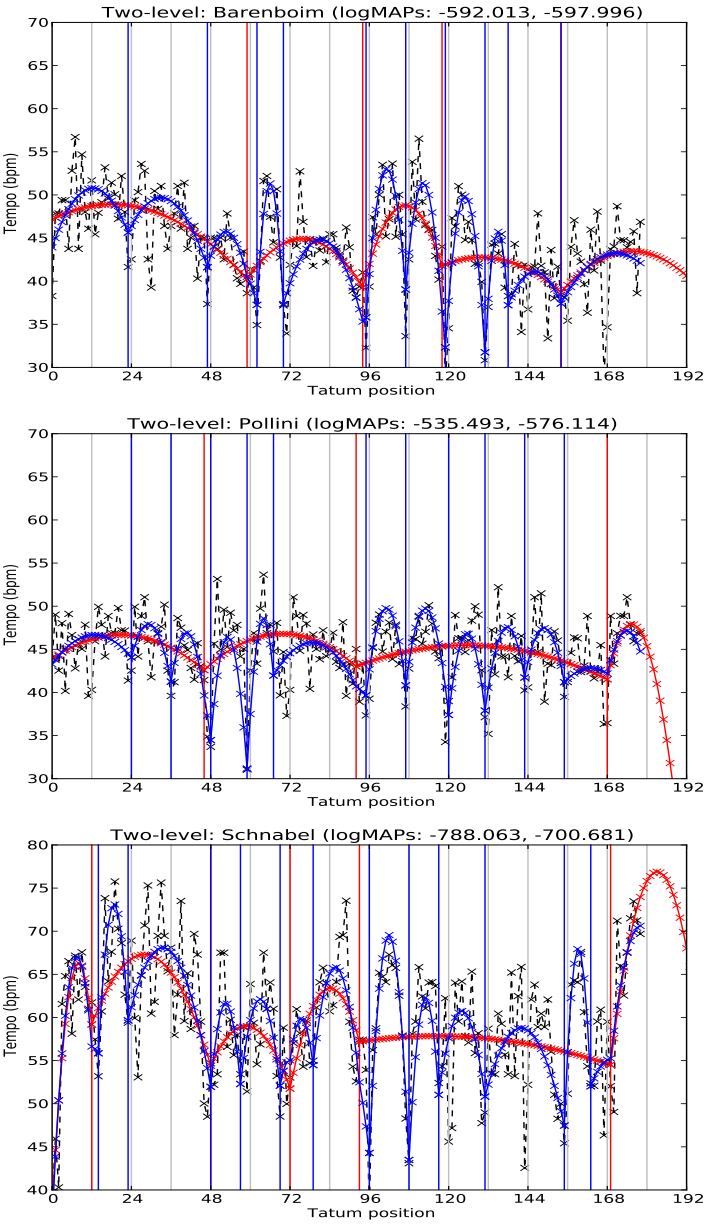


Fig. 2. Two-level analysis of performances by each of three pianists (Barenboim, Pollini, Schnabel). In each plot, the first long-scale fit (centred on the four-bar timescale) is depicted in red, and the second shorter-scale fit (centred on the one-bar timescale) is given in blue. The second fit is pre-offset by the first, meaning the blue arcs display the combined model produced by both timescales combined. Annotated data finish at tatum 180; where the MAP choice extends beyond that, we show the predicted immediate future arc.

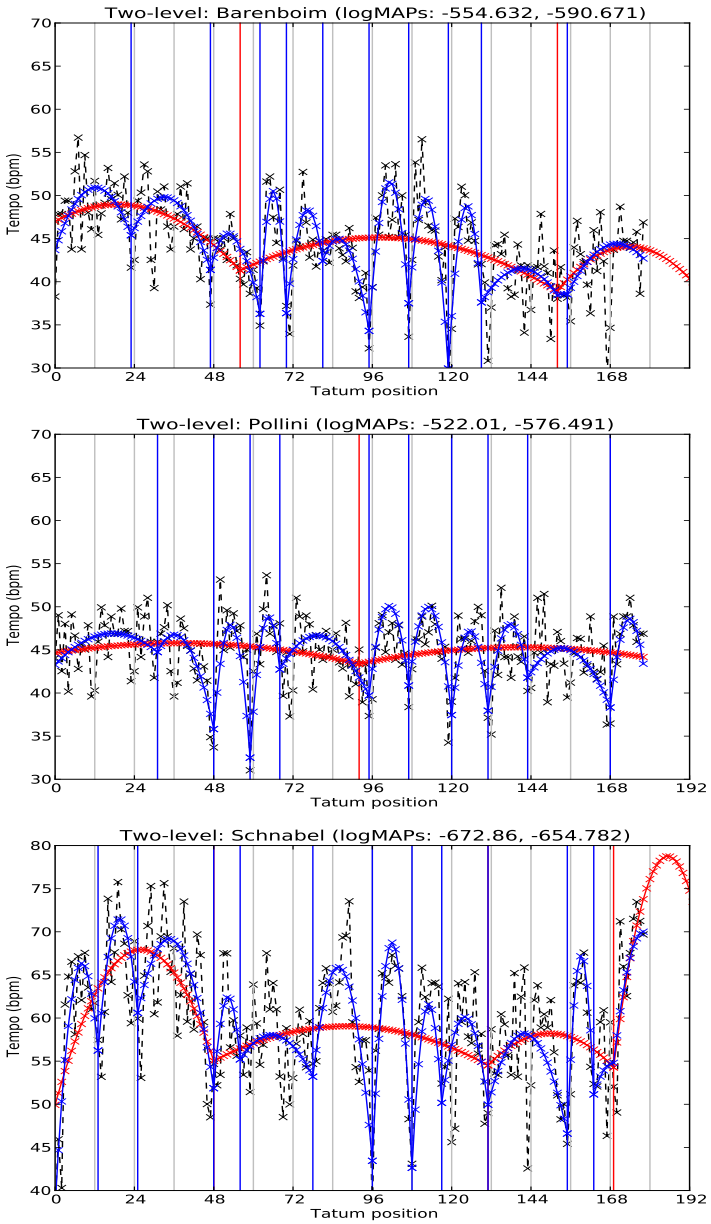


Fig. 3. As Figure 2 but with the standard deviation of the noise prior set at 4.0 rather than 3.0.

References

[1] P. E. Allen and R. B. Dannenberg. Tracking musical beats in real time. In *Proc. International Computer Music Conference (ICMC)*, pages 140–143, Hong Kong, 1990.

[2] C. M. Bishop. *Pattern Recognition and Machine Learning*, volume 4. Springer, New York, 2006.

[3] E. Cheng and E. Chew. Quantitative analysis of phrasing strategies in expressive performance: computational methods and analysis of performances of unaccompanied bach for solo violin. *Journal of New Music Research*, 37(4):325–338, 2008.

[4] E. Chew. About time: Strategies of performance revealed in graphs. *Visions of Research in Music Education*, 20(1), Jan 2012.

[5] C. H. Chuan and E. Chew. A dynamic programming approach to the extraction of phrase boundaries from tempo variations in expressive performances. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 305–308, Vienna, Austria, 2007.

[6] R. B. Dannenberg and S. Mohan. Characterizing tempo change in musical performances. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 650–656, 2011.

[7] N. P. McAngus Todd. The dynamics of dynamics: A model of musical expression. *Journal of the Acoustical Society of America*, 91(6), 1992.

[8] A. Robertson and M. D. Plumbley. B-Keeper: A beat-tracker for live performance. In *Proc. International Conference on New Interfaces for Musical Expression (NIME), New York, USA*, pages 234–237, 2007.

[9] G. Widmer and A. Tobudic. Playing Mozart by analogy: Learning multi-level timing and dynamics strategies. *Journal of New Music Research*, 32(3):259–268, 2003.

Structural Similarity Based on Time-span Tree

Satoshi Tojo[1] and Keiji Hirata[2]

[1] Japan Advanced Institute of Science and Technology tojo@jaist.ac.jp
[2] Future University Hakodate hirata@fun.ac.jp

Abstract. Time-span tree is a stable and consistent representation of musical structure since most experienced listeners deliver the same one, almost independently from context and subjectivity. In this paper, we pay attention to the reduction hypothesis of the tree structure, and introduce the notion of distance as a promising candidate of stable and consistent metric of similarity. First, we design a feature structure to represent a time-span tree. Next, we regard that when a branch is removed from the tree the information corresponding to its time-span is lost, and suggest that the sum of the length of those removed spans is the distance between two trees. We will show that the distance preserves uniqueness in multiple shortest paths, as well as the triangle inequality. Thereafter, we illustrate how the distance works as a metric of similarity, and then, we discuss the feasibility and the problem of our methodology.

Keywords: Similarity, time-span reduction, feature structure, join, meet

1 Introduction

As is remarked in [26], *an ability to assess similarity lies close to the core of cognition*. Musical similarity is multi-faceted as well [15], and this property inevitably raises a context-dependent, subjective behavior [14]. As to context dependency, similarity cannot be perceived in isolation from the musical context in which it occurs. Volk stated in [22]: *Depending on the context, similarity can be described using very different features*. For instance, the impact of cultural knowledge may degrade a stable similarity assessment. As to subjectivity, similarity is likely perceived differently between subjects and even within a subject, depending on listening style, preference, and so on. For instance, [23] revealed that the inconsistency in the annotations by experts is caused by the divergence of four musical dimensions (rhythm, contour, motif, and mode).

Thus far, many researches have explored stable and consistent musical similarity metrics as a central topic in music modelling and music information retrieval [9, 4]. Some of them are motivated by engineering demands such as musical retrieval, classification, and recommendation [15, 7, 18], and others are by modelling the cognitive processes of musical similarity [5, 6]. In this paper, we also seek for a stable and consistent similarity, postponing context-dependency and subjectivity later. We regard that similarity is stable in the sense that similarity assessment is performed only on a score of music, disregarding such context-dependent factors as timber, artist, subject matter of lyrics, and cultural factors. Also, we regard that similarity assessment is consistent in the sense that most experienced listeners can deliver same results as long as the western-tonal-classical style of music is targeted.

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

To propose a stable and consistent similarity, we rely on the cognitive reality or perceptual universality of music theory. As addressed in [24], *systems which aim to encode musical similarity must do so in a human-like way*. Now, we take the stance that *tree* structure underlies such cognitive reality. Bod claimed in his DOP model [1] that there lies cognitive plausibility in combining a rule-based system with a fragment memory when a listener parses music and produces a relevant tree structure, like a linguistic model. Lerdahl and Jackendoff presumed that perceived musical structure is internally represented in the form of hierarchies, which means time-span tree and strong reduction hypothesis in Generative Theory of Tonal Music (GTTM, hereafter) [16, p.2, pp.105-112, p.332]. Dibben argued that the experimental results show that pitch events in tonal music are heard in a strict hierarchical manner and provide evidence for the internal cognitive representation of time-span tree of GTTM [3]. Wiggins et al. deployed discussions on the tree structures and argued that they are more about semantic grouping than about syntactic grouping [25]. We basically follow their view, under which we assume the time-span tree of a melody represents its meaning. Here, we need to admit that GTTM has its innate problem, that is, those ambiguous preference rules may result in multiple time-span analyses; [8] has solved this issue, assigning a parametric weight to each rule, and has implemented an automatic tree analyzer.

In effect, tree representation has contributed to the study on similarity. Marsden began with conventional tree representations and allowed joining of branches in the limited circumstances with preserving the directed acyclic graph (DAG) property for expressing information dependency [13]. As a result, high expressiveness was achieved, while it was difficult to define consistent similarity between melodies. Valero proposed a representation method dedicated to a similarity comparison task, called metrical tree [21]. Valero used a binary tree representing the metrical hierarchy of music and avoided the necessity of explicitly encoding onsets and duration; only pitches needed to be encoded. As a measure to compare metrical trees, Valero adopted the tree edit distance with many parameters, which were justified only by the best performance in experiments, but not by cognitive reality.

Among the properties of time-span tree, in particular, we consider the concept of *reduction* essential, when a time-span tree subsumes a reduced one. Selfridge-Field also claimed that a relevant way of taking deep structures (meaning) into account is to adopt the concept of reduction [19]. Since the subsumption relation between time-span trees can be defined as a partial order, the above consideration may imply a possibility for treating time-span tree (i.e., the meaning of a melody) as a mathematical entity. Our objective is to derive the notion of distance from the reduction and the subsumption relation, to employ it as a metric of similarity. At this time, our attitude toward the design is strictly computational; that is, there must lie a reliable logical and algebraic structure so that we will be able to implement the similarity onto computers.

In the following Section 2, we translate a time-span tree into a feature structure, carefully preventing the other factors from slipping into the structure, to guarantee stability. In Section 3, we define a notion of distance between time-span trees and then show that the notion enjoys several desirable mathematical properties, including the triangle inequality. In Section 4, we illustrate our analysis. In Section 5 we discuss open

problems concerning how we can apply our notion of distance to music similarity, and in Section 6 we summarize our contribution.

2 Time-Span Tree in Feature Structure

In this section, we develop the representation method for time-span tree in [11, 10], in terms of feature structure. First we introduce the general notion of feature structure, and then we propose a set of necessary features to represent a time-span tree. As the set of feature structures are partially ordered, we define such algebraic operations as *meet* and *join* and show that the set becomes a *lattice*. Since this section and the following section include mathematical foundation, those who would like to see examples first may jump to Section 4 and come back to technical details afterward.

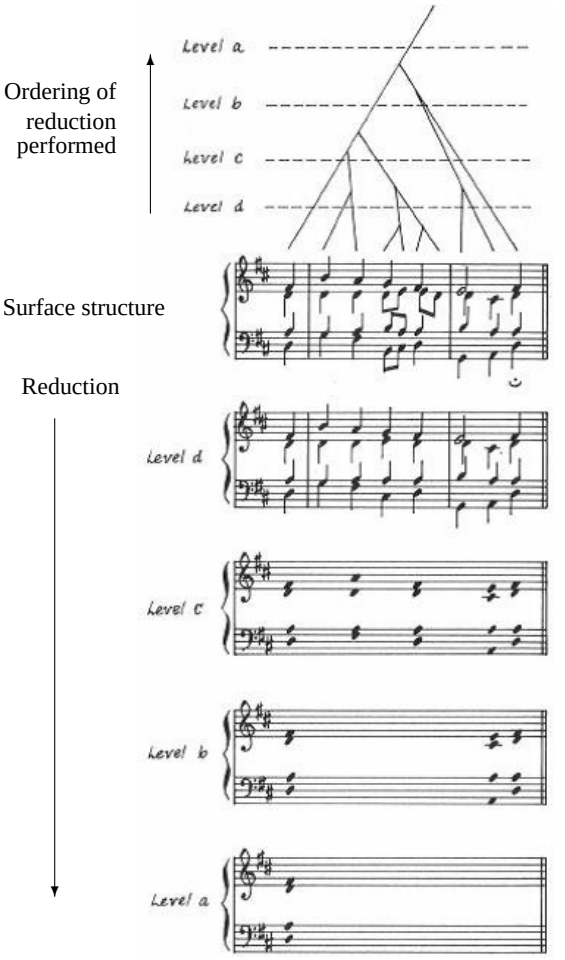


Fig. 1. Time-span reduction in GTTM (Lerdahl and Jackendoff [16, page 115])

2.1 Time-Span Tree and Reduction

A melody is considered to be a sequence of pitch events in temporal order, consisting of a single note and a chord. Time-span reduction [16] assigns structural importance to each pitch events in the hierarchical way. The structural importance is derived from the grouping analysis, in which multiple notes compose a short phrase called a group, and from the metrical analysis, where the regular alternation of strong and weak beats affects. As a result, a time-span tree becomes a binary tree constructed in bottom-up and top-down manners by comparison between the structural importance of adjacent pitch events at different hierarchical levels.

Fig. 1 shows an excerpt from [16] demonstrating the concept of reduction. In the sequence of reductions, each level should sound like a natural simplification of the previous level.[3] The alternative omission of notes must make the successive levels sound less like the original. Hence, reduction can be regarded as rewriting an expression to an equivalent simpler one; it often has the same meaning as abstraction. Since reduction is designed based on Gestalt grouping, the reduction successfully associates a melody with another one that sounds quite similar. The key idea of our framework is that reduction is identified with the subsumption relation, which is the most fundamental relation in knowledge representation.

2.2 Feature Structure and Subsumption Relation

Feature structure (*f-structure*, hereafter) has been mainly studied for applications to linguistic formalism based on unification and constraint, such as Head-driven Phrase Structure Grammar (HPSG)[2, 17]. An f-structure is a list of feature-value pairs where a value may be replaced by another f-structure recursively. Below is an f-structure in attribute-value matrix (AVM) notation where σ is a structure, the label headed by '~ ' (tilde) is the *type* of the whole structure, and f_i's are feature labels and v_i's are their values:

$$\sigma = \begin{bmatrix} \tilde{}type \\ f_1\ v_1 \\ f_2\ v_2 \end{bmatrix} .$$

A type requires its indispensable features. When all these intrinsic features are properly valued, the f-structure is said to be *full-fledged*.

Now we define the notion of *subsumption*. Let σ_1 and σ_2 be f-structures. σ_2 subsumes σ_1, that is, $\sigma_1 \sqsubseteq \sigma_2$ if and only if for any $(f\ \ v_1) \in \sigma_1$ there exists $(f\ \ v_2) \in \sigma_2$ and $v_1 \sqsubseteq v_2$. Since we suppose an f-structure is considered to be the conjunctive set of feature-value pairs, '$\sqsubseteq$' corresponds to the so-called Hoare order of sets (e.g., $\{b, d\} \sqsubseteq \{a, b, c, d\}$). For example, by assuming $v_1 \sqsubseteq [f_3\ \ v_3]$, σ_1 below is subsumed both by the following σ_2 and σ_3.

$$\sigma_1 = \begin{bmatrix} \tilde{}type1 \\ f_1\ v_1 \end{bmatrix}, \qquad \sigma_2 = \begin{bmatrix} \tilde{}type1 \\ f_1\ v_1 \\ f_2\ v_2 \end{bmatrix}, \qquad \sigma_3 = \begin{bmatrix} \tilde{}type1 \\ f_1 \begin{bmatrix} \tilde{}type2 \\ f_3\ v_3 \end{bmatrix} \end{bmatrix} .$$

[3] Once a melody is reduced, each note with onset and duration properties becomes a virtual note that is just a pitch event dominating a corresponding time-span, omitting onset and duration. Therefore, to listen to a reduced melody, we assume that it can be rendered by regarding a time-span as a real note with such onset timing and duration.

Since both σ_2 and σ_3 are elaborations of σ_1, which are differently elaborated, ordering '$\sqsubseteq$' is a partial order, not a total order like integers and real numbers. Equivalence $a = b$ is defined as $a \sqsubseteq b \wedge b \sqsubseteq a$.

To denote value v of feature f in structure σ, we write $\sigma.f = v$. Thus, $\sigma_1.f_1 = v_1$ and $\sigma_1.f_2$ is undefined while $\sigma_3.f_1.f_3 = v_3$. We call a sequence of features $f_1.f_2.\cdots.f_n$ a *feature path*. Structure sharing is indicated by boxed tags such as $\boxed{\text{i}}$ or $\boxed{\text{j}}$. The set value $\{x, y\}$ means the choice either of x or y, and $\bot$ means that the value is empty. Even for $\bot$, any feature f_i is accessible though $\bot.f_i = \bot$.

2.3 Time-Span Trees in F-Structures

We name the type of an f-structure corresponding a time-span tree $\tilde{}tree$.

Definition 1 (Tree Type F-structure) *A full-fledged $\tilde{}tree$ f-structure possesses the following features.*

- *head represents the most salient pitch event in the tree.*
- *span represents the length of the time-span of the whole tree, measured by the number of quarter notes.*
- *dtrs (daughters) are subtrees, whose left and right are recursively $\tilde{}tree$. This dtrs feature is characterized by the following two conditions.*
  - *The value of span must be the addition of two spans of the daughters.*
  - *The value of $head$ is chosen from either that of $left$ or of $right$ daughter.*

If $head = dtrs.left.head$, the node has the right-hand elaboration of shape $\diagup\!\!\diagdown$, while if $head = dtrs.right.head$, the left-hand elaboration $\diagdown\!\!\diagup$. If $dtrs = \bot$ then the tree consists of a single branch with a single pitch event at its leaf.

Fig. 2 shows the examples. Such bold-face letters as **C4**, **E4** and **G4** are pitch events.

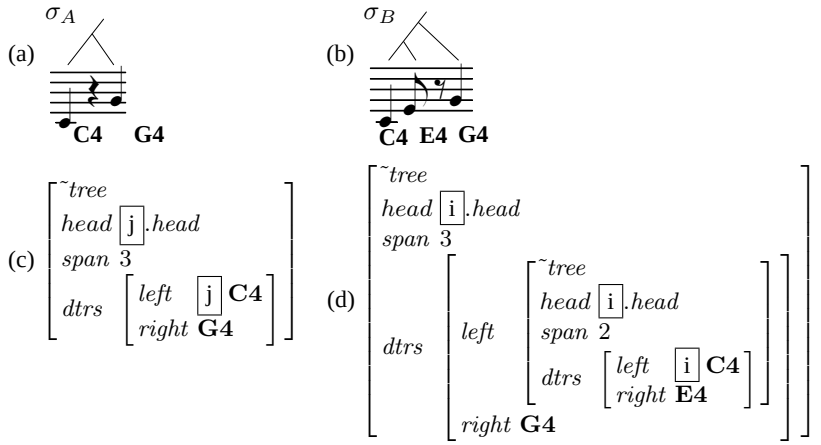


Fig. 2. Melodies (a) and (b) and their f-structures (c) and (d), respectively.

The value of *head* feature is occupied by $\tilde{}event$ f-structure; a full-fledged one should include *pitch*, *onset*, and *duration* features. For example,

$$\mathbf{C4} = \begin{bmatrix} \tilde{}tree & \\ head & \begin{bmatrix} \tilde{}event & \\ pitch & C4 \\ onset & \ldots \\ duration & 1 \end{bmatrix} \\ span & \ldots \\ dtrs & \bot \end{bmatrix}.$$

2.4 Unification, *Join* and *Meet*

Intuitively, unification is a process of information conjunction. We introduce the set notation of an f-structure using the set of feature-path-value pairs: $\{(f_{11}.\cdots.f_{1n}\ v_1), (f_{21}.\cdots.f_{2m}\ v_2), \cdots\}$. Unification is the consistent union of f-structures in the set notation, results in another f-structure. Unification fails only if there exists an inconsistency in any feature-path-value pair.

The set of f-structures are partially ordered as there is the subsumption relation. Here, we can introduce *join* and *meet* operations; *Join* corresponds to a union of sets or a consistent overlay while *meet* does to intersection or the common part.

Definition 2 (*Join*) *Let σ_A and σ_B be full-fledged f-structures representing the time-span trees of melodies A and B, respectively. If we can fix the least upper bound of σ_A and σ_B, that is, the least y such that $\sigma_A \sqsubseteq y$ and $\sigma_B \sqsubseteq y$ is unique, we call such y the join of σ_A and σ_B, denoted as $\sigma_A \sqcup \sigma_B$.*

Theorem 3.13 in Carpenter [2] provides that the unification of f-structures A and B is the least upper bound of A and B, which is equivalent to *join* in this paper. Similarly, we regard the intersection of the unifiable f-structures as *meet*.

Definition 3 (*Meet*) *Let σ_A and σ_B be full-fledged f-structures representing the time-span trees of melodies A and B, respectively. If we can fix the greatest lower bound of σ_A and σ_B, that is, the greatest x such that $x \sqsubseteq \sigma_A$ and $x \sqsubseteq \sigma_B$ is unique, we call such x the meet of σ_A and σ_B, denoted as $\sigma_A \sqcap \sigma_B$.*

We show a musical example in Fig. 3.

Obviously from Definitions 2 and 3, we obtain the absorption laws: $\sigma_A \sqcup x = \sigma_A$ and $\sigma_A \sqcap x = x$ if $x \sqsubseteq \sigma_A$. Moreover, if $\sigma_A \sqsubseteq \sigma_B$, for any x $x \sqcup \sigma_A \sqsubseteq x \sqcup \sigma_B$ and $x \sqcap \sigma_A \sqsubseteq x \sqcap \sigma_B$.

We can define $\sigma_A \sqcup \sigma_B$ and $\sigma_A \sqcap \sigma_B$ in recursive functions. In the process of unification between σ_A and σ_B, when we are to match a subtree with a single branch in the counterpart, if we always choose the subtree the result becomes $\sigma_A \sqcup \sigma_B$ and if we always choose the single branch we obtain $\sigma_A \sqcap \sigma_B$. Because there is no alternative action in these procedures, $\sigma_A \sqcup \sigma_B$ and $\sigma_A \sqcap \sigma_B$ exist uniquely. Thus, the partially ordered set of time-span trees becomes a *lattice*.

Since time-span tree T is rigidly corresponds to f-structure σ, we identify T with σ and may call σ a tree in the following sections as long as no confusion.

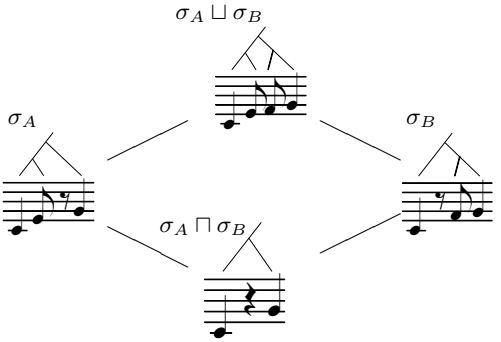


Fig. 3. *Join* and *Meet* operations of time-span trees

3 Strict Distance in Time-Span Reduction

In this section, we introduce the notion of distance between two time-span trees. We propose that:

> *If a branch with a single pitch event is reduced, the information corresponding to the length of its time-span is lost.*

Thus, we regard the accumulation of such lost time-spans as the distance of two trees in the sequence of reductions, called *reduction path*. Thereafter, we generalize the notion to be feasible, not only in a reduction path but in any direction in the lattice. Finally in this section, we show the distance suffices the triangle inequality. Again as this section includes technical details, those who would like to see examples earlier may skip this section and can come back later.

We restrict that branches are reduced only one by one, for the convenience to sum up distances. A branch is *reducible* only when there exists no other lower branch than its junction (attaching point); thus, a reducible branch possesses a single pitch event at its leaf. In the similar way, we restrict that a branch can be an elaboration of some tree only when it consists of a single event and can be attached to a junction under which there is no other branch.

By the way, the *head* pitch event of a tree structure is the representative of the whole tree, whose length appears at *span* feature. Though the event itself retains its original shorter duration, we may regard its supremacy is extended to the tree length. The situation is the same as each subtree. Thus, we consider that each pitch event has the maximal length of dominance.

Definition 4 (Maximal Time-span) *Each pitch event has the maximal time-span within which the event becomes most salient, and outside the time-span its supremacy is lost.*

In Fig. 4, a reducible branch on pitch event e_2 has the time-span s_2. After e_2 is reduced, branch on e_1 becomes reducible and the connected span $s_1 + s_2$ becomes e_1's maximal time-span, though its original duration was s_1. Finally, after e_1 is reduced, e_3 dominates the length of $s_1 + s_2 + s_3$. When e_2 and e_1 are reduced in this order, the distance between σ_A and σ_C becomes $s_2 + (s_1 + s_2)$.

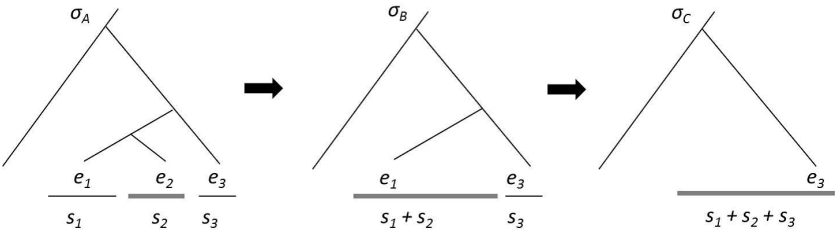


Fig. 4. Reduction by maximal time-spans; gray thick lines denote maximal time-spans while thin ones pitch durations.

Prior to the formal definition of distance, we impose *Head/Span Equality Condition* (*HSEC*, hereafter):

$$\sigma_A.head = \sigma_B.head \ \ \& \ \ \sigma_A.span = \sigma_B.span.$$

We have included this restriction in the following algorithm, so as to avoid any futile comparison; if the identity of two heads and their time-spans is disregarded, the distance between them is meaningless.

Let $\varsigma(\sigma)$ be a set of pitch events in σ, $\sharp\varsigma(\sigma)$ be its cardinality, and s_e be the maximal time-span of event e. Since reduction is made by one reducible branch at a time, a reduction path $\sigma_B = \sigma^n, \sigma^{n-1}, \ldots, \sigma^2, \sigma^1, \sigma^0 = \sigma_A$ suffices $\sharp\varsigma(\sigma^{i+1}) = \sharp\varsigma(\sigma^i) + 1$. For each reduction step, when a reducible branch on event e disappears, its maximal time-span s_e is accumulated as distance.

Definition 5 (Reduction Distance) *The distance $d_{\sqsubseteq}$ of two time-span trees such that $\sigma_A \sqsubseteq \sigma_B$ in a reduction path is defined by*

$$d_{\sqsubseteq}(\sigma_A, \sigma_B) = \sum_{e \in \varsigma(\sigma_B) \setminus \varsigma(\sigma_A)} s_e.$$

Although the distance is a simple summation of maximal time-spans at a glance, there is a latent order in adding the spans, for reducible branches change dynamically in the process of reduction. In order to give a constructive procedure on this summation, we introduce the notion of total sum of maximal time-spans.

Definition 6 (Total Maximal Time-span) *Given ˜tree f-structure σ,*

$$tms(\sigma) = \sum_{e \in \varsigma(\sigma)} s_e.$$

We present $tms(\sigma)$ as a recursive function in Algorithm 1.

Input: a ˜*tree* f-structure σ
Output: $tms(\sigma)$
1 **if** $\sigma = \bot$ **then**
2 **return** 0;
3 **else if** $\sigma.dtrs = \bot$ **then**
4 **return** $\sigma.span$;
5 **else**
6 **case** $\sigma.head = \sigma.dtrs.left.head$
7 **return** $tms(\sigma.dtrs.left) + tms(\sigma.dtrs.right) + \sigma.dtrs.right.span$;
8 **case** $\sigma.head = \sigma.dtrs.right.head$
9 **return** $tms(\sigma.dtrs.left) + tms(\sigma.dtrs.right) + \sigma.dtrs.left.span$;

Algorithm 1: Total Maximal Time-span

In Algorithm 1, Lines 1–2 are the terminal condition. Lines 3–4 treat the case that a tree consists of a single branch. In Lines 6–7, when the right subtree surrender to the left, the left extends the domination rightward by $\sigma.dtrs.right.span$. Ditto for the case the right-hand side overcomes the left, as Lines 8–9.

When $\sigma_A \sqsubseteq \sigma_B$, from Definition 5 and 6,

$$\begin{aligned} d_{\sqsubseteq}(\sigma_A, \sigma_B) &= \textstyle\sum_{e \in \varsigma(\sigma_B) \setminus \varsigma(\sigma_A)} s_e = \sum_{e \in \varsigma(\sigma_B)} s_e - \sum_{e \in \varsigma(\sigma_A)} s_e \\ &= tms(\sigma_B) - tms(\sigma_A). \end{aligned}$$

As a special case of the above, $d_{\sqsubseteq}(\bot, \sigma) = tms(\sigma)$.

Next, we consider the notion of distance that can be applicable to two trees reside in different paths.

Lemma 1 *For any reduction path from $\sigma_A \sqcup \sigma_B$ to $\sigma_A \sqcap \sigma_B$, $d_{\sqsubseteq}(\sigma_A \sqcap \sigma_B, \sigma_A \sqcup \sigma_B)$ is unique.*

Proof As there is a reduction path between $\sigma_A \sqcap \sigma_B$ and $\sigma_A \sqcup \sigma_B$, and $\sigma_A \sqcap \sigma_B \sqsubseteq \sigma_A \sqcup \sigma_B$, $d_{\sqsubseteq}(\sigma_A \sqcap \sigma_B, \sigma_A \sqcup \sigma_B)$ is computed by the difference of total maximal time-span in Algorithm 1. Because the algorithm returns a unique value, the distance is unique. ∎

Theorem 1 (Uniqueness of Reduction Distance) *If there exist reduction paths from σ_A to σ_B, $d_{\sqsubseteq}(\sigma_A, \sigma_B)$ is unique.*

Lemma 2 $d_{\sqsubseteq}(\sigma_A, \sigma_A \sqcup \sigma_B) = d_{\sqsubseteq}(\sigma_A \sqcap \sigma_B, \sigma_B)$ *and* $d_{\sqsubseteq}(\sigma_B, \sigma_A \sqcup \sigma_B) = d_{\sqsubseteq}(\sigma_A \sqcap \sigma_B, \sigma_A)$.

Proof From set-theoretical calculus, $\varsigma(\sigma_A \sqcup \sigma_B) \setminus \varsigma(\sigma_A) = \varsigma(\sigma_A) \cup \varsigma(\sigma_B) \setminus \varsigma(\sigma_A) = \varsigma(\sigma_B) \setminus \varsigma(\sigma_A) \cap \varsigma(\sigma_B) = \varsigma(\sigma_B) \setminus \varsigma(\sigma_A \sqcap \sigma_B)$. Then, by Definition 5, $d_{\sqsubseteq}(\sigma_A, \sigma_A \sqcup \sigma_B) = \sum_{e \in \varsigma(\sigma_A \sqcup \sigma_B) \setminus \varsigma(\sigma_A)} s_e = \sum_{e \in \varsigma(\sigma_B) \setminus \varsigma(\sigma_A \sqcap \sigma_B)} s_e = d_{\sqsubseteq}(\sigma_A \sqcap \sigma_B, \sigma_B)$. ∎

Definition 7 (Meet and Join Distances)

– $d_{\sqcap}(\sigma_A, \sigma_B) = d_{\sqsubseteq}(\sigma_A \sqcap \sigma_B, \sigma_A) + d_{\sqsubseteq}(\sigma_A \sqcap \sigma_B, \sigma_B)$ *(meet distance)*

– $d_{\sqcup}(\sigma_A, \sigma_B) = d_{\sqsubseteq}(\sigma_A, \sigma_A \sqcup \sigma_B) + d_{\sqsubseteq}(\sigma_B, \sigma_A \sqcup \sigma_B)$ *(join distance)*

Lemma 3 $d_{\sqcup}(\sigma_A, \sigma_B) = d_{\sqcap}(\sigma_A, \sigma_B)$.

Proof Immediately from Lemma 2. ∎

Lemma 4 *For any* σ', σ'' *such that* $\sigma_A \sqsubseteq \sigma' \sqsubseteq \sigma_A \sqcup \sigma_B$, $\sigma_B \sqsubseteq \sigma'' \sqsubseteq \sigma_A \sqcup \sigma_B, d_{\sqcup}(\sigma_A, \sigma') + d_{\sqcap}(\sigma', \sigma'') + d_{\sqcup}(\sigma'', \sigma_B) = d_{\sqcup}(\sigma_A, \sigma_B)$. *Ditto for the meet distance.*

Now the notion of distance, which was initially defined in the reduction path as $d_{\sqsubseteq}$ is now generalized to $d_{\{\sqcap,\sqcup\}}$, and in addition we have shown they have the same values. From now on, we omit $\{\sqcap, \sqcup\}$ from $d_{\{\sqcap,\sqcup\}}$, simply denoting 'd'.

Theorem 2 (Uniqueness of Distance) $d(\sigma_A, \sigma_B)$ *is unique among shortest paths between* σ_A *and* σ_B.

Note that shortest paths can be found in ordinary graph-search methods, such as *branch and bound*, Dijkstra's algorithm, best-first search, and so on.

Corollary 1 $d(\sigma_A, \sigma_B) = d(\sigma_A \sqcup \sigma_B, \sigma_A \sqcap \sigma_B)$.

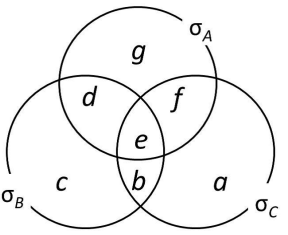
Proof From Lemma 2 and Lemma 3. ∎

Theorem 3 (Triangle Inequality) *For any* σ_A, σ_B *and* σ_C, $d(\sigma_A, \sigma_B) + d(\sigma_B, \sigma_C) \geq d(\sigma_A, \sigma_C)$.

Proof From Corollary 1 and by definition,

$$d(\sigma_i, \sigma_j) = d(\sigma_i \sqcup \sigma_j, \sigma_i \sqcap \sigma_j) = \textstyle\sum_{e \in \varsigma(\sigma_i \sqcup \sigma_j) \setminus \varsigma(\sigma_i \sqcap \sigma_j)} s_e.$$

Since we employ the set-notation of f-structure (cf. Section 2.4), the relationship between $\sigma_{\{A,B,C\}}$ can be depicted in Venn diagram. Then, $d(\sigma_A, \sigma_B) + d(\sigma_B, \sigma_C)$ becomes the sum of maximal time-spans in $\varsigma(\sigma_A \sqcup \sigma_B) \setminus \varsigma(\sigma_A \sqcap \sigma_B)$ plus those in $\varsigma(\sigma_B \sqcup \sigma_C) \setminus \varsigma(\sigma_B \sqcap \sigma_C)$, which corresponds to $(f + g + b + c) + (a + c + d + f) = a + b + 2c + 2f + d + g$ in the diagram. On the contrary, $d(\sigma_A, \sigma_C)$ becomes the sum of $a + b + d + g$. Since $(a+b+2c+2f+d+g) - (a+b+d+g) = 2c+2f \geq 0$, we obtain the result. ∎



In the above proof, c and f are counted twice because branches in these areas are once reduced and later added, or once added and later reduced. This implies that these reduction/addition can be skipped and there exists a short cut between σ_A and σ_C without visiting σ_B.

Finally in this section, we suggest that the distance can be a metric of similarity between two music pieces. As long as we stay in the lattice of reductions under *HSEC*, the distance exactly reflects the similarity. However, even though *head*s and *span*s are different in two pieces of music, we can calculate the similarity with our notion of distance. We show such examples in Section 4.

4 Examples

In this section, we illustrate our analyses. The first example is Mozart's K265, *Ah! vous dirais-je, maman*, equivalent to *Twinkle, Twinkle, Little Star*. The melody in the left-hand side of Fig. 5 is the theme, while those in the right-hand side are the third variation and its reduced melodies in downward order. The horizontal lines below each score are the maximal time-spans of pitch events though we omit explicit connection between events and lines in the figure. The lines drawn at the bottom level in each score correspond to reducible branches (i.e., reducible pitch events) at that step. For example, from Level c in the right-hand side of Fig. 5 to Level b, eight maximal time-spans of $1/3$-long disappear by reduction, thus, according to Algorithm 1 the distance is $1/3 \times 8 = 8/3$. The configuration of maximal time-spans at Level a in the right-hand

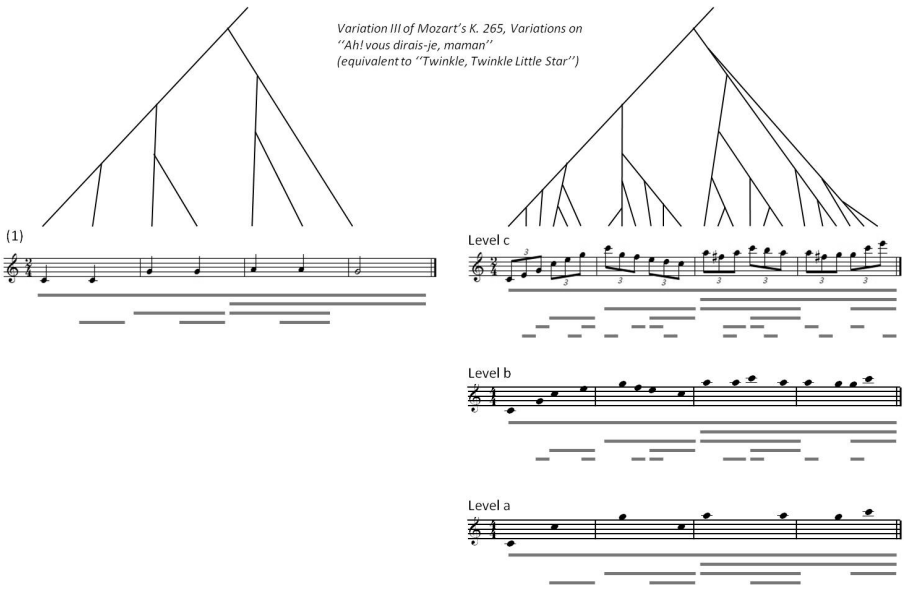


Fig. 5. Reduction of *Mozart: Ah! vous dirais-je, maman*

of Fig. 5 quite resembles that in the left-hand side, which is the theme of the variation. Actually, since the difference between (1) and Level a is the rightmost quarter note in the 4-th measure, the distance between these two is so close as just 1. This implies that we can retrieve the theme by reducing the variation.

In Fig. 6, we have arranged various reductions originated from a piece. As we can find three reducible branches in A we possess three different reductions: B, C, and D. In the figure, C (shown diluted) lies at the back of the lattice where three back-side edges meet.

The distances, represented by the length of edges, from A to B, D to F, C to E, and G to H are same, since the reduced branch is common. Namely, the reduction

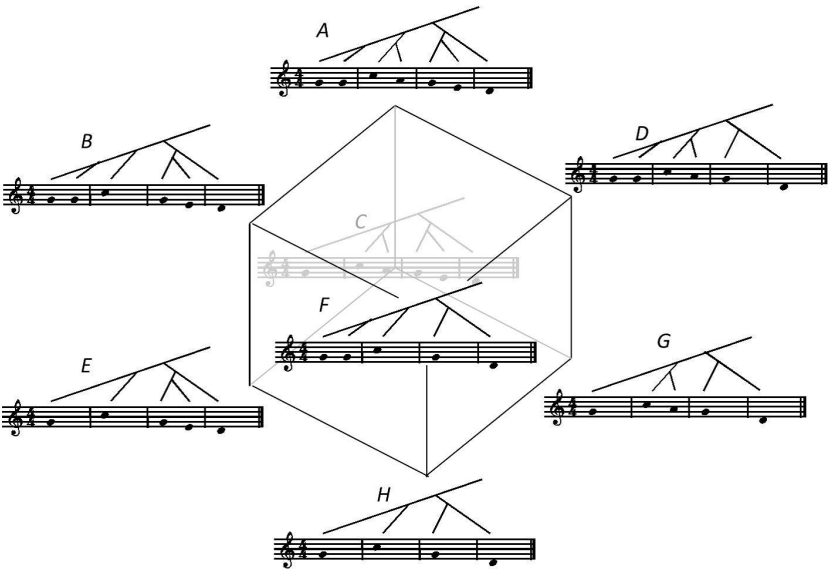


Fig. 6. Reduction lattice

lattice becomes parallelepiped,[4] and the distances from A to H becomes uniquely $2 + 2 + 2 = 6$, which we have shown as Theorem 1. We exemplify the triangle inequality (Theorem 3); from A through B to F, the distance becomes $2+2 = 4$, and that from F through D to G is $2+2 = 4$, thus the total path length becomes $4+4 = 8$. But, we can find a shorter path from A to G via D, in which case the distance becomes $2 + 2 = 4$. Notice that the lattice represents the operations of *join* and *meet*; e.g., $F = B \sqcap D$, $D = F \sqcup G$, $H = E \sqcap F$, and so on. In addition, the lattice is locally Boolean, being A and H regarded to be $\top$ and $\bot$, respectively. That is, there exists a complement,[5] and $E^c = D$, $C^c = F$, $B^c = G$, and so on.

In the next example, we compare two time-span trees in reduction. The left-hand side in Fig. 7 is *Massa's in De Cold Ground* (Stephen Collins Foster, 1852) and the right-hand side is *Londonderry Air* (transposed to C major). The vertical distance is strictly computable in each reduction, but in addition, we may notice that these two pieces are quite near in their skeletons in the abstract levels. Especially, we should compare the configurations of maximal time-spans in the bottom three levels and find them topologically equal to each other. This means the distance becomes 0, being *HSEC* disregarded. Then, in the next section, we discuss how to compute the distance where *HSEC* does not hold.

[4] In the case of Fig. 6, as all the edges have the length of 2, the lattice becomes a cube.
[5] For any member X of a set, there exists X^c and $X \sqcup X^c = \top$ and $X \sqcap X^c = \bot$.

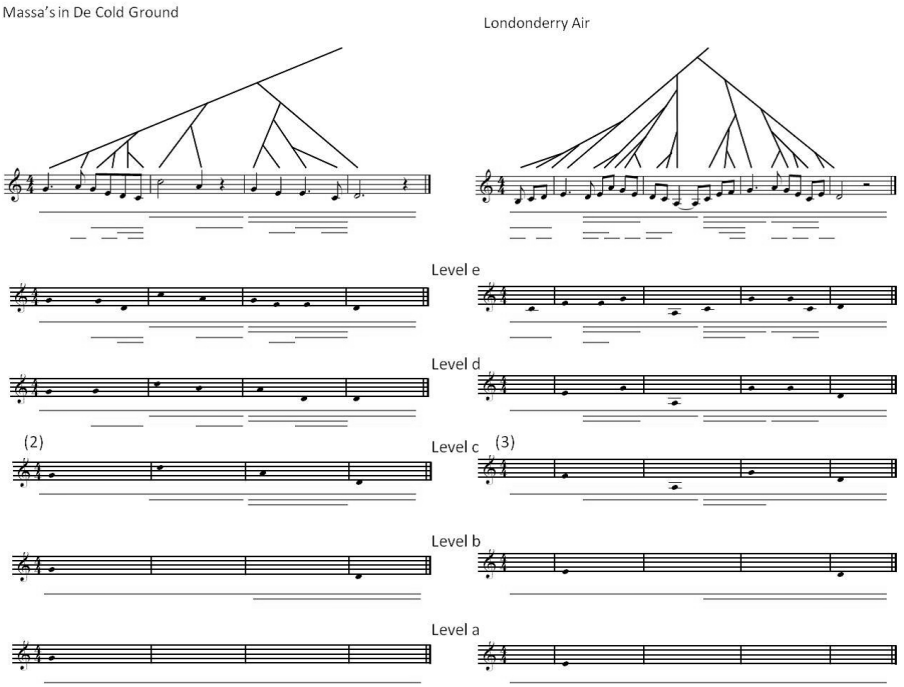


Fig. 7. Reduction processes of *Massa's in De Cold Ground* and *Londonderry Air*

5 Discussion

In this section, we discuss several open problems. In Section 2, we have introduced the representation of time-span tree in f-structure and *join* and *meet* operations, which however only work properly under *HSEC*. From a practical point of view, this condition is too restrictive for arbitrarily given two melodies. We found that *Massa's in De*

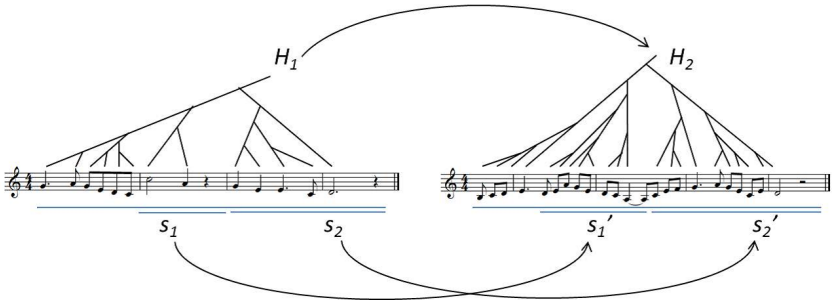


Fig. 8. flexible matching

Cold Ground and *Londonderry Air* do not share strictly common time-span trees, but are somewhat similar as a result of reduction as in Fig. 7. Since we actually recognize a flavor of similarity in them, we have a good reason to seek for a more flexible mechanism to map *head*s and *span*s as in Fig. 8 in *join* and *meet* computation. The situation is same for the comprison of pitch events residing at *head* feature. For the purpose, we have to provide the subsumption relations in time-spans and in pitch events, grounded to cognitive reality; if these partial orders truly coincide with our intuition or perception, we can tolerate the condition of unificaiton.

The similarity measures widely used in data mining and information retrieval include Jaccard, Simpson, Dice, and Point-wise mutual information (PMI) [20]. For instance, the Jaccard index (also known as Jaccard similarity coefficient) is regarded as an index of the similarity of two sets.

$$sim(\sigma_A, \sigma_B) = \frac{|\sigma_A \sqcap \sigma_B|}{|\sigma_A \sqcup \sigma_B|},$$

Here, we may naïvely interpret '$|\sigma|$' as the set of pitch events in the tree as '$\sharp_\varsigma(\sigma)$'. However, the number of notes does not fully reflect the internal structure. Then, it may be appropriate to weight an individual note by its time-span, and the content of a structure hence amounts to the total maximal time-span $tms(\sigma)$ in Definition 6, as

$$sim(\sigma_A, \sigma_B) = \frac{tms(\sigma_A \sqcap \sigma_B)}{tms(\sigma_A \sqcup \sigma_B)}.$$

Since the value of $tms(\sigma)$ represents the complexity of the whole structure, we can also consider the *density* of notes in the music piece. Similarly, we may make use of Simpson index with tms as follows:

$$sim(\sigma_A, \sigma_B) = \frac{tms(\sigma_A \sqcap \sigma_B)}{min(tms(\sigma_A), tms(\sigma_B))}.$$

We have treated the maximal time-spans evenly, independent of their lengths and levels at which they occur. However, suppose we listen to two melodies of the same length; one is with full of short notes while the other with a few long notes, then the psychological lengths of these two melodies may be different. This effect is actually well known as the Weber-Fechner law; the relationship between stimulus and perception is logarithmic in auditory and visual psychology. Since our initial purpose of this paper has been to present a stable and consistent similarity, we could not reflect such perceptional aspects.

6 Conclusions

In this paper, we relied on the strong reduction hypothesis of the tree structure in GTTM, and presented the notion of metric of similarity, based on the distance of reduction. In order to do that, we first designed an f-structure to represent a time-span tree, and we showed that its *head* feature and *span* feature properly reflected the original structure

proposed in GTTM. Thereafter, we regarded that a reduction was the loss of information, and the loss was quantified by the time-span of a reduced event. We defined the notion of distance by the lost time-span, and have generalized the notion as the metric of similarity. We have shown several mathematical properties concerning the metric, including uniqueness of distance in any shortest paths as well as the triangle inequality.

Our contribution in this paper is two-fold. One is that we have presented a stable and consistent metric of similarity, which does rely on neither subjective nor context-dependent factor. The other is that our metric is mathematically so sound that it can be employed in the framework of well-known traditional measures, such as Jaccard/Simpson indices.

At present, we have the following five open problems entangled each other. First, (i) if we are to apply our unification mechanism such as *join* and *meet* operations to practical problems, e.g., melodic morphing, we need to ease *HSEC*. Also, (ii) we need more statistical witness in comparison of such existing metrics as Jaccard/Simpson indices, referring to a large-scale music database. As was mentioned in Section 5, (iii) we have treated the maximal time-spans evenly, disregarding the psychological length of music. Since we have postponed such subjective and context-dependent metric, we are obliged to face this aspect from now. By the way, (iv) we still have various alternatives to render each reduced event on actual staff. Though we have mentioned this in the footnote 3 in Section 2.1, the problem is left undone. Finally, (v) the more fundamental problem is the reliability of time-span tree. We admit that some processes in the time-span reduction is still fragile and proper reduction is not promised yet. Thus far we have tackled the automatic reduction system, and even from now on we need to improve the system performance. All in all, to apply such an objective metric to practical cases we need further consideration, that would be our future works.

Acknowledgment

The authors would like to thank the all anonymous reviewers for their fruitful comments, which helped us to develop the contents and to improve the readability. This work was supported by KAKENHI 23500145, Grants-in-Aid for Scientific Research of JSPS.

References

1. Bod, R.: A Unified Model of Structural Organization in Language and Music. Journal of Artificial Intelligence Research 17, 289–308 (2002)
2. Carpenter, B.: The Logic of Typed Feature Structures. Cambridge University Press (1992)
3. Dibben, N.: Cognitive Reality of Hierarchic Structure in Tonal and Atonal Music. Music Perception 12(1), 1–25 (Fall 1994)
4. Downie, J.S., Byrd, D., Crawford, T.: Ten Years of ISMIR: Reflections of Challenges and Opportunities. In: Proceedings of ISMIR 2009, 13–18
5. ESCOM: 2007 Discussion Forum 4A. Similarity Perception in Listening to Music. Musicæ Scientiæ
6. ESCOM: 2009 Discussion Forum 4B. Musical Similarity. Musicæ Scientiæ

7. Grachten, M., Arcos, J.-L., de Mantaras, R.L.: Melody retrieval using the Implication/Realization model. 2005 MIREX. http://www.music-ir.org/evaluation/mirexresults/articles/similarity/grachten.pdf
8. Hamanaka, M., Hirata, K., Tojo, S.: Implementing "A Generative Theory of Tonal Music". Journal of New Music Research 35(4), 249–277 (2007)
9. Hewlett, W.B., Selfridge-Field, E.: Melodic Similarity. Computing in Musicology 11, The MIT Press (1998)
10. Hirata, K., Tojo, S.: Lattice for Musical Structures and Its Arithmetics. LNAI 4384 (Selected Papers from JSAI 2006, T. Washio et al. (Eds)) Springer-Verlag, 54–64 (2007)
11. Hirata, K., Tojo, S., Hamanaka, M.: Melodic Morphing Algorithm in Formalism, In: Proceedings of 3rd International Conference, MCM 2011 (LNAI 6726), 338–341
12. Lartillot, O.: Multi-Dimensional Motivic Pattern Extraction Founded on Adaptive Redundancy Filtering. Journal of New Music Research 34(4), 375–393 (2005)
13. Marsden, A.: Generative Structural Representation of Tonal Music. Journal of New Music Research 34(4), 409–428 (2005)
14. Ockelford, A.: Similarity relations between groups of notes: Music-theoretical and music-psychological perspectives. In: Musicae Scientiae, Discussion Forum 4B, Musical Similarity, 47–98 (2009)
15. Pampalk, E.: Computational Models of Music Similarity and their Application in Music Information Retrieval. PhD Thesis, Vienna University of Technology (March 2006)
16. Lerdahl, F., Jackendoff, R.: A Generative Theory of Tonal Music. The MIT Press (1983)
17. Sag, I.A., Wasow, T.: Syntactic Theory: A Formal Introduction. CSLI Publications (1999)
18. Schedl, M., Knees, P., Böck, S.: Investigating the Similarity Space of Music Artists on the Micro-Blogosphere. In: Proceedings of ISMIR 2011, 323–328
19. Selfridge-Field, E.: Conceptual and Representational Issues in Melodic Comparison. Computing in Musicology 11, 3–64 (1998)
20. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining. Addison-Wesley (2005)
21. Valero, D.R.: Symbolic Music Comparison with Tree Data Structure. Ph.D. Thesis, Universitat d' Alacant, Departamento de Lenguajes y Sistemas Informatícos (2010)
22. Volk, A., Wiering, F.: Music Similarity. In: ISMIR 2011 Tutorial on Musicology. http://ismir2011.ismir.net/tutorials/ISMIR2011-Tutorial-Musicology.pdf
23. Volk, A., van Kranenburg, P., Garbers, J., Wiering, F., Veltkamp, R.C., Grijp, L.P.: A manual annotation method for melodic similarity and the study of melody feature sets. In: Proceedings of ISMIR 2008, 101–106
24. Wiggins, G.A.: Semantic Gap?? Schematic Schmap!! Methodological Considerations in the Scientific Study of Music. In: 2009 11th IEEE International Symposium on Multimedia, 477–482
25. Wiggins, G.A., Müllensiefen, D., Pearce, M.T.: On the non-existence of music: Why music theory is a figment of the imagination. In: Musicae Scientiae, Discussion Forum 5, 231–255 (2010)
26. Wilson, R.A., Keil, F. (Eds): The MIT Encyclopedia of the Cognitive Sciences. The MIT Press (May 1999)

Subject and counter-subject detection for analysis of the Well-Tempered Clavier fugues

Mathieu Giraud[1], Richard Groult[2], and Florence Levé[2]

[1] LIFL, CNRS, Université Lille 1 and INRIA Lille, France
[2] MIS, Université Picardie Jules Verne, Amiens, France

Abstract. Fugue analysis is a challenging problem. We propose an algorithm that detects subjects and counter-subjects in a symbolic score where all the voices are separated, determining the precise ends and the occurrence positions of these patterns. The algorithm is based on a diatonic similarity between pitch intervals combined with a strict length matching for all notes, except for the first and the last one. On the 24 fugues of the first book of Bach's *Well-Tempered Clavier*, the algorithm predicts 66% of the subjects with a musically relevant end, and finally retrieves 85% of the subject occurrences, with almost no false positive.

Keywords: symbolic music analysis, contrapuntal music, fugue analysis, repeating patterns

1 Introduction

Contrapuntal music is a polyphonic music where each individual line bears interest in its own. Bach fugues are a particularly consistent model of contrapuntal music. The fugues of Bach's *Well-Tempered Clavier* are composed of two to five voices, appearing successively, each of these voices sharing the same initial melodic material: a subject and, in most cases, a counter-subject. These patterns, played completely during the exposition, are then repeated all along the piece, either in their initial form or more often altered or transposed, building a complex harmonic network.

To analyze symbolic scores with contrapuntal music, one can use generic tools detecting repeating patterns or themes, possibly with approximate occurrences. Similarity between a pattern and several parts of a piece may be computed by the Mongeau-Sankoff algorithm [20] and its extensions or by other methods for approximate string matching [7, 8], allowing a given number of restricted mismatches. Several studies focus on finding *maximal repeating patterns*, limiting the search to *non-trivial* repeating patterns, that is discarding patterns that are a sub-pattern of a larger one with the same frequency [13, 14, 16, 17]. Other studies try to find musically significant *themes*, with algorithms considering the number of occurrences [25], but also the melodic contour or other features [18].

Some MIR studies already focused on contrapuntal music. The study [26] builds a tool to decide if a piece is a fugue or not, but no details are given on the

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

algorithm. The bachelor thesis [1] contains a first approach to analyze fugues, including voice separation. For sequence analysis, it proposes several heuristics to help the selection of repeating patterns inside the algorithms of [13] which maximizes the number of occurrences. The website [10] also produces an analysis of fugues, extracting sequences of some repeating patterns, but without precise formal analysis nor precise bounds.

One can take advantage of the apparently simple structure of a fugue: as the main theme – the subject – always begins at only one voice, this helps the analysis. But a good understanding of the fugue requires to find *where the subject exactly ends*. In this work, we start from a symbolic score which is already track-separated, and we propose an algorithm to sketch the plan of the fugue. The algorithm tries to retrieve the *subjects* and the *counter-subjects*, precisely determining the *ends* of such patterns. We tested several substitution functions to have a sensible and specific approximate matching. Our best results use a simple *diatonic similarity* between pitch intervals [4] combined with a strict length matching for all notes, except for the first and the last one.

The paper is organized as follows. Section 2 gives definitions and some background on fugues, Section 3 details the problem of the bounds of such patterns, Section 4 presents our algorithm, and Section 5 details the results on the 24 fugues of the first book of Bach's *Well-Tempered Clavier*. These results were evaluated against a reference musicological book [2]. The algorithm predicts two thirds of the subjects with a musically relevant end, and finally retrieves 85% of the subject occurrences, with almost no false positives.

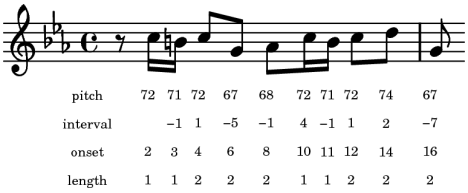
2 Preliminaries

A *note* x is described by a triplet (p, o, ℓ), where p is the pitch, o the onset, and ℓ the length. The pitches can describe diatonic (based on note names) or semitone information. We consider ordered *sequence of notes* $x_1 \dots x_m$, that is $x_1 = (p_1, o_1, \ell_1), \dots, x_m = (p_m, o_m, \ell_m)$, where $1 \leq o_1 \leq o_2 \leq \dots \leq o_m$ (see Fig. 1). The sequence is *monophonic* if there are never two notes sounding at the same onset, that is, for every i with $1 \leq i < m$, $o_i + \ell_i \leq o_{i+1}$. To be able to match transposed patterns, we consider relative pitches, also called *intervals*: the interval sequence is defined as ${}^{\Delta}x_2 \dots {}^{\Delta}x_m$, where ${}^{\Delta}x_i = ({}^{\Delta}p_i, o_i, \ell_i)$ and ${}^{\Delta}p_i = p_i - p_{i-1}$.

We now introduce some notions about fugue analysis (see for example [2, 23] for a complete musicological analysis). These concepts are illustrated by an example on Fugue #2, which has a very regular construction.

A *fugue* is given by a set of *voices*, where each voice is a monophonic sequence of notes. In Bach's *Well-Tempered Clavier*, the fugues have between 2 and 5 voices, and Fugue #2 is made of 3 voices.

The fugue is built on a theme called *subject* (S). The three first *occurrences* of the subject in Fugue #2 are detailed in Fig. 2: the subject is *exposed* at one voice



pitch	72	71	72	67	68	72	71	72	74	67
interval		−1	1	−5	−1	4	−1	1	2	−7
onset	2	3	4	6	8	10	11	12	14	16
length	1	1	2	2	2	1	1	2	2	2

Fig. 1. A monophonic sequence of notes (start of Fugue #2, see Fig. 2), represented by (p, o, ℓ) or $(\triangle p, o, \ell)$ triplets. In this example, onsets and lengths are counted in sixteenths, and pitches and intervals are counted in semitones through the MIDI standard.

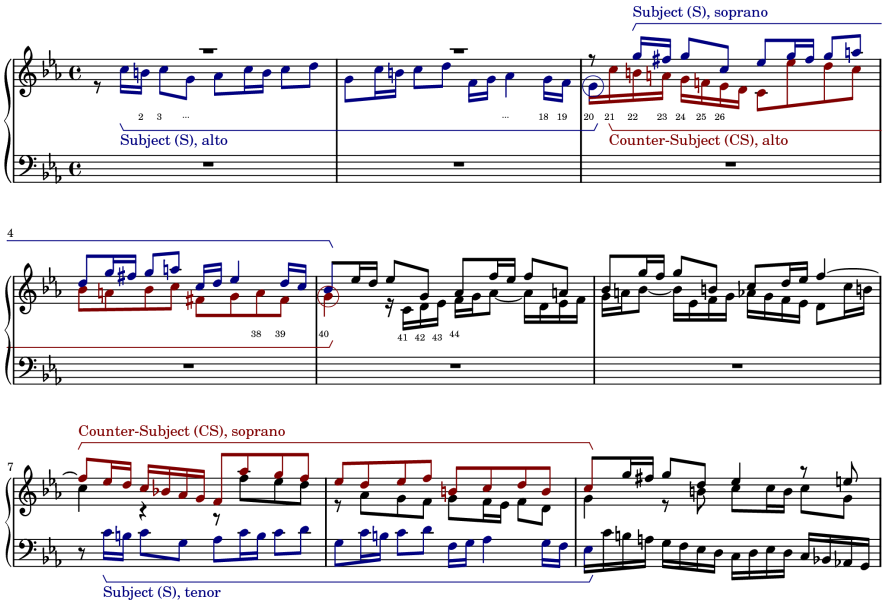
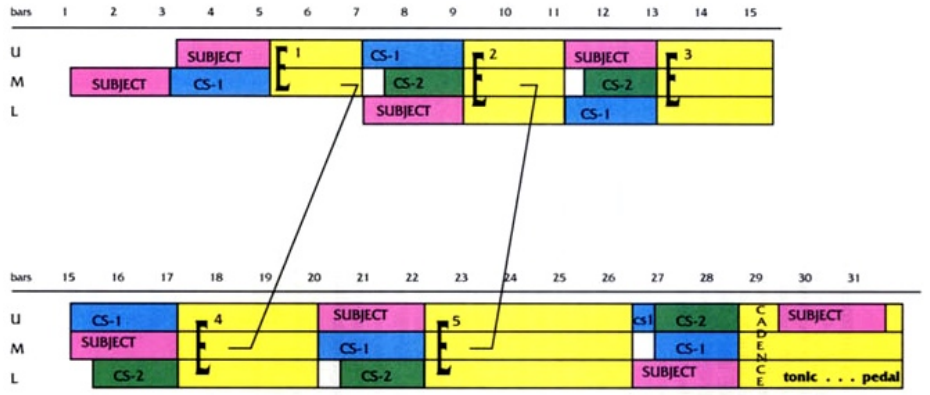


Fig. 2. Start of Fugue #2 in C minor (BWV 847).

(the alto), beginning by a C, until the second voice enters (the soprano, measure 3). The subject is then exposed at the second voice, but is now transposed to G. Meanwhile, the first voice continues with the first *counter-subject* (CS) that combines with the subject. Fig. 3 shows a sketch of the entire fugue. The fugue alternates between other instances of the subject together with counter-subjects (8 instances of S, 6 instances of CS, and 5 instances of the *second counter-subject* CS2) and development on these same patterns called *episodes* (E).

All these instances are not exact ones – the patterns can be transposed or altered in various ways. As an example, Fig. 4 shows the five complete occurrences of CS. For these occurrences, the patterns can be (diatonically) transposed, and the lengths are conserved except for the first and last note.




```
             1   2   3   4   5   6   7   8   9  10   1   2   3   4
soprano              S17-----       C19-----         S19-----
   alto      S19-----CS19----
  tenor                                S19-----         C19-----
             I       V                  I                III

5   6   7   8   9  20   1   2   3   4   5   6   7   8   9  30   1
CS18----           S19-----                                S19-----
S17-----           CS19----
                                           S19-----
V                  I                       I           I
```

Fig. 3. Analysis of Fugue #2 in C minor (BWV 847). Top: diagram summarizing the analysis by S. Bruhn, used with permission [2], [3, p. 80]. Bottom: output of the proposed algorithm, retrieving all occurrences of S (and their degrees) and all but one occurrences of CS. The numbers indicate the pitch intervals exactly matching (in a diatonic way) those of the patterns (out of 19 for S). The two `S17` occurrences correspond thus to approximate matches of the subject (tonal answers).

3 Where does the subject end?

A fundamental question concerns the precise *length* of the subject and of any other interesting pattern. The subject is heard alone at the beginning of the first voice, until the second voice enters. However, this end is generally not exactly at the start of the second voice.

Formally, let us suppose that the first voice is $x_1, x_2, ...$, and the second one is $y_1, y_2, ...$, with $x_i = (p_i, o_i, \ell_i)$ and $y_j = (p'_j, o'_j, \ell'_j)$. Let x_z be the last note of the first voice heard before or at the start of the second voice, that is $z = \max\{i \mid o_i \leq o'_1\}$. The end of the subject is roughly at x_z. Table 2, at the end of the paper, lists the exact values of g such that the *true* subject ends at x_{z+g}:

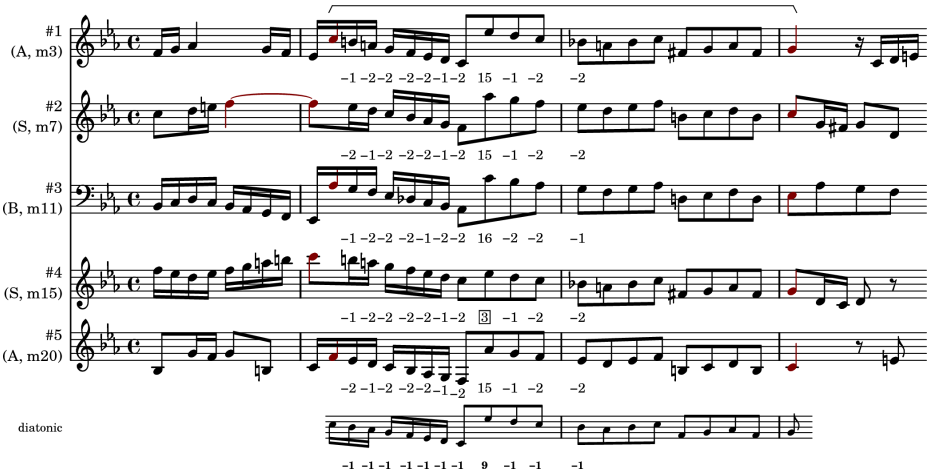


Fig. 4. The 5 complete occurrences of the first counter-subject into Fugue #2 in C minor (BWV 847). (Note that this counter-subject actually has a latter occurrence, split between two voices.) In these occurrences, all notes – except the first and the last ones – have exactly the same length. The values in the occurrences indicate the intervals, in number of semitones, inside the counter-subject. Only occurrences #2 and #5 have exactly the same intervals. The occurrence #4 is almost identical to occurrence #1, except that it lacks the octave jump (+3 instead of +15). Between groups {#1, #4}, {#2, #5}, and {#3}, the intervals are not exactly the same. However, all these intervals (except the lack of the octave jump in #4) are equal when one considers only diatonic information (bottom small staff): clef, key and alterations are here deliberately omitted, as semitone information is not considered.

in the first book of Bach's *Well-Tempered Clavier*, we notice that g is always between -8 and $+6$, and, in the majority of cases, between -4 and $+1$.

For example, in the Fugue #2, the subject has 20 notes, ending on alto note x_{20} (the first sixteenth of the third measure, E♭, first circled note on Fig. 2), that is 2 notes before the start of the soprano voice ($g = -2$). This can be deduced from many observations in this third measure:

- metrically, the phrase ends on a strong beat;
- harmonically, the five preceding notes "F G A♭ G F" suggest a 9th dominant chord, which resolves on the E♭ suggesting the C minor tonic;
- moreover, the subject ends with a succession of sixteenths with small intervals, whereas the following note x_{21} (C) belongs to CS with the line of falling sixteenths.

Determining the precise end of the subject is thus an essential step in the analysis of the fugue: it will help to localize the counter-subject and build the structure with all occurrences of these patterns, but also to understand the rhythm, the harmony and the phraseology of the whole piece.

We could use generic algorithms to predict the subject end. For example, the "stream segment detection" described in [24] considers melody, pitch and rhythm informations. Many different features are also discussed in [18] for theme extraction. However, in the following, we will show that a simple algorithm, only based on similarities, is able to detect precisely most of the subject ends.

4 Algorithm

Starting from track-separated data, we propose here to detect the subject as a repeating pattern finishing approximatively at the start of the second voice, under a substitution function considering a diatonic similarity for pitch intervals, and enforcing length equalities of all notes except the first one and the last one.

The similarity score between a pattern and the rest of the fugue piece can be computed via dynamic programming by the Mongeau-Sankoff equation [20]. The alignment can then be retrieved through backtracking in the dynamic programming table.

As almost all the content of a fugue is somewhat derived from a subject or some counter-subject, any part will match a part of the subject or of another base pattern within a given threshold. Here, we will use very conservative settings – only substitution errors, and strict length requirements – to have as few false positives as possible, still keeping a high recognition rate.

Subject identification. To precisely find the end the subject, we thus want to test patterns finishing at notes x_{z+g}, where $g \in [g_{\min}, g_{\max}] = [8, +6]$. Each one of these candidates is matched against all the voices. In this process, we use a substitution cost function able to match the first and the last notes of the subject independently of their lengths.

Let $S(a, b)$ be the best number of matched intervals when aligning the start $x_1 \dots x_a$ of a pattern (the subject) against a part of a given voice finishing at y_b, and $S_f(a, b)$ the best number of matched intervals when aligning a complete pattern (the complete candidate subject) $x_1 \dots x_a$ against the same part. These tables S and S_f may be computed by the following dynamic programming equation:

$$\begin{cases} S(1, b) = 0 \\ \forall a \geq 2, S(a, b) = S(a-1, b-1) + \delta({}^{\Delta}x_a, {}^{\Delta}y_b) & \text{(match, substitution)} \\ \forall a \geq 2, S_f(a, b) = S(a-1, b-1) + \delta_f({}^{\Delta}x_a, {}^{\Delta}y_b) & \text{(finishing)} \end{cases}$$

The substitution functions δ and δ_f are detailed on Fig. 5: δ checks pitch intervals and lengths, whereas δ_f only considers pitch intervals. The length of the first notes (x_1 and y_1) is neither checked, as the algorithm actually compares ${}^{\Delta}x_2 \dots {}^{\Delta}x_a$ against ${}^{\Delta}y_2 \dots {}^{\Delta}y_b$. Finally, notice that these equations only use substitution operations, but can be extended to consider other edit operations.

$$\delta((^{\Delta}p, o, \ell), (^{\Delta}p', o', \ell')) = \begin{cases} +1 & \text{if} \quad {}^{\Delta}p \approx {}^{\Delta}p' \quad \text{and} \quad \ell = \ell' \\ 0 & \text{if} \quad {}^{\Delta}p \not\approx {}^{\Delta}p' \quad \text{and} \quad \ell = \ell' \\ -\infty & \text{otherwise} \end{cases}$$

$$\delta_f((^{\Delta}p, o, \ell), (^{\Delta}p', o', \ell')) = \begin{cases} +1 & \text{if} \quad {}^{\Delta}p \approx {}^{\Delta}p' \\ 0 & \text{otherwise} \end{cases}$$

Fig. 5. Substitution operations between intervals. The actual comparison of length ($\ell = \ell'$) also checks the equality of the rests that may be immediately before the compared notes. The relation $\approx$ is a similarity relation on pitch intervals.

As in [13], we only compute once each table (for a given voice), then we scan the table S_f to find the occurrences: given a sequence x and a threshold τ, the candidate finishing at x_{z+g} occurs in the sequence y if for some position i in the text, $S_f(z+g, i) \geq \tau$. The best candidate is selected on the total number of matched intervals in all occurrences. We call x_s its last note, so the subject is defined to be $x_1 ... x_s$. The whole algorithm is in $O(mn)$, where $m = z + g_{\max}$.

For example, on the Fugue #2, the algorithm correctly selects the note x_{20} as the end of the subject (see Table 1).

$z+g$	14	15	16	17	18	19	**20**	21	22	23	24	25	26	27	28
g	-8	-7	-6	-5	-4	-3	$\mathbf{-2}$	-1	0	$+1$	$+2$	$+3$	$+4$	$+5$	$+6$
occ.	8	8	8	8	8	8	**8**	3	3	3	3	2	2	2	2
score	100	108	116	124	132	140	**148**	59	61	63	65	48	50	52	54

Table 1. Occurrences and scores when matching all candidate subjects in Fugue #2. The score is the sum of the $S_f(z+g, i)$ values at least equal to τ: it is the total number of intervals exactly matched on all occurrences. Here this end corresponds to the "non-trivial maximal-length repeating pattern" for most occurrences, but it is not always the case.

Interval similarities and diatonic matching. The equation Fig. 5 needs to have a similarity relation $\approx$ on pitch intervals. Between a strict pitch equality and very relaxed "up/down" classes defining the contour of some melody [9], some intermediary interval classes may be defined as "step/leap intervals" [5] or "quantized partially overlapping intervals" (QPI) [15].

We propose here to use a similarity on *diatonic pitches*. Such a pitch representation is often mentioned [21, 22] and was studied in [4, 6, 12]. A diatonic model is very relevant for tonal music: it is sensible enough to allow mode changes, while remaining specific enough – a scale will always match only a scale. For example, with diatonic similarity, all occurrences but one of the counter-subject on the Fig. 4 can be retrieved exactly, and the occurrence #4 with only one substitution.

Counter-subjects identification. The same method as for subject identification is used to retrieve the first counter-subject. The first occurrence of the counter-subject starts right after the subject (at x_{s+1}), and its length is approximatively equal to the length of the subject. We thus have a rough end of the counter-subject at x_w, where $w = \max\{i \mid o_i - o_{s+1} \leq o_s - o_1\}$, and the same procedure refines the bound to find the good last note x_{cs}, where again $cs - w$ is in a given interval. To prevent detection of non-relevant patterns, the counter-subject is marked as not detected if the above procedure leads to more occurrences than the subject occurrences.

5 Results and discussion

We tested the algorithm of the previous section on the 24 fugues of the first book of Bach's *Well-Tempered Clavier*, starting from Humdrum files where the voices are separated, available for academic purposes at `http://kern.humdrum.org/`. The pitches were encoded according to two frameworks: MIDI encoding, and Base40 encoding [11]. While the first one only counts semitones, the second one allows to discriminate enharmonic pitches, thus allowing a precise diatonic match as described in the previous section.

We ran the algorithm on the 24 fugues[3], and manually checked all results and occurrences. Results (with diatonic similarity) are summarized on Table 2. We fixed a minimum threshold of $\tau = 0.9z - 3$, where z is the number of notes defined in Section 3.

Subject lengths. We searched for end of subjects in the range $[g_{\min}, g_{\max}] = [-8, +6]$, that are the observed values. In 16 of the 24 fugues, the algorithm retrieves precisely the ends of the subjects. To our knowledge, this is the first algorithm able to correctly detect the ends of the subject: In [1], the subjects found are said to be "missing or including an extra 1 to 4 notes", and the ends of the subjects on [10] are also very approximate.

Fugue #8 shows why the proposed algorithm does not always find the correct length of the subject. In this fugue, a subject of length 9 notes is found instead of 13 notes: there are several truncated occurrences of the subject, and the algorithm chooses the end that provides the best match throughout the piece (Fig. 6).

The algorithm already considers the last note in a special way (and the former notes can be handled through substitution errors in the pitch intervals). It is possible to adapt the matching to be even more relaxed towards the end of the pattern, but we did not see a global improvement in the detection of subject lengths.

False positives. There are very few false positives among the subjects found (specificity of 90%), even when the length of the subject is badly predicted. The false positives appear in only two fugues:

[3] A part of the output of the algorithm is shown at the bottom of Fig. 3, and the full output on all the 24 fugues is available at `http://www.lifl.fr/~giraud/fugues`.

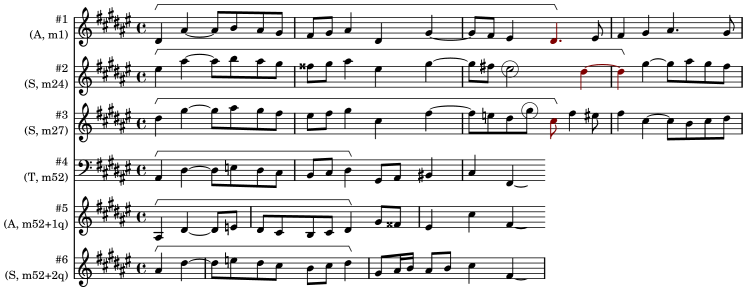


Fig. 6. Some subject occurrences in Fugue #8 in D# minor. The occurrence #1 is the first one, and is similar to 16 occurrences, sometimes with diatonic transpositions. In the occurrence #2, the last but one note of the subject (circled E) has not the same length than in the other occurrences (and this is forbidden by our substitution function δ). In the occurrence #3, a supplementary note (circled G) is inserted before the end of the subject, again preventing the detection if the true length of the subject is considered. Moreover, the occurrences #3, #4 and #5 are truncated to the head of the subject, and lead to a false detection of subject length.

- in Fugue #19, the 5 false positives correspond to 4 extended subjects [2], and one almost complete subject.
- in Fugue #5, the length of the subject is wrongly selected to the first 9 notes (8 first thirty-second notes and a final note), and this head of the subject matches the 11 true occurences, but also 24 false positives.

False negatives. The algorithm correctly retrieves about 85% of the subject occurrences. The false negatives are occurrences that are too much altered: insertions, deletions, or too many substitutions compared to the threshold.

Inverted and augmented subjects. In some fugues, the subject appears upside down (all its intervals are reversed) or augmented (all lengths are doubled). Once the subject is known, the same matching algorithm can thus be applied to the inversion or the augmentation of the subject. This method never produced a false positive, and was able to recover 72% (26/36) of the complete inverted and augmented subjects reported in [2].

Counter-subjects. Counter-subjects were detected with the same algorithm within the range $[g_{\min}, g_{\max}] = [-2, +4]$. In 40% of the fugues, the algorithm correctly detects the exact length of the CS or the absence of a CS.

In 9 fugues, the algorithm predicts the absence of CS. This was expected for Fugues #1, #8 (no CS), #15 (the CS occurs completely only once) #19 (late exposition of CS) and #20 (there is no real "characteristic and independent counter-subject" according to [2]). As in the case of the subjects, there are false negatives due to the bad recogntion of altered patterns. Moreover, when the subject is badly detected, the detection of the counter-subject end fails in the majority of the cases.

The algorithm retrieves correctly about the half of the CS occurrences, with more than 80% specificity.

Pitch interval similarities. We compared the diatonic matching against a simple exact matching on MIDI semitones, possibly adapting the error threshold. As expected, diatonic similarity has a better performance, because such a relaxed similarity is able to match approximate occurrences as the counter-subjects shown on Fig. 4.

Starting from MIDI pitches, an idea could be thus to use pitch spelling methods as [19]: such methods are almost perfect and provide the diatonic spelling of some pitches. However, we also tested a pseudo-diatonic matching on semitone information – considering as similar the intervals that differ from at most 1 semitone. The results (not shown) are very similar to those with true diatonic matching.

Other edit operations. Finally, we also tested other edit operations. The equations of Fig. 5 consider only substitutions, and can be simply extended to include the full Mongeau-Sankoff edit operations [20]. For instance, using insertions and allowing rhythm substitutions will, starting from the true subject, retrieve the occurrences #2 and #3 in Fig. 6. However, in the general case, insertions or deletions destroy the measure, leading to bad results on the predicted subject lengths.

More musical operations (fragmentation, consolidation), with fine-tuned costs, give a slight advantage in some of the 24 fugues, but this has not been reported here to keep the simplicity of the algorithm.

6 Conclusions

A complete fugue analysis tool should use any available information, including pattern repetition, harmonic analysis and phrasing considerations.

In this work, we focused only on pattern repetition. Our simple algorithm, based on the total number of matched intervals in all occurrences of patterns, allows to find precise ends of subjects and first counter-subjects in the majority of cases. This model considers a unique substitution operation with a diatonic similarity, enforcing the equality of lengths for all notes except the first and the last ones.

Extensions could include a study on the second counter-subject and on other inferred patterns. Combined with other techniques, this algorithm could lead to a more complete automatic fugue analysis pipeline.

Track-separated data. The current algorithm works on track-separated data. Starting from plain MIDI files, we could use voice separating algorithms. Although it would be a challenging problem to adapt our algorithm to directly treat standard polyphonic MIDI files, we first want to improve the current approach to complete our comprehension of any fugue.

Studying other fugues. Finally, it would be interesting to study the efficiency of our algorithm on other fugues than Bach's Well-Tempered Clavier, keeping in mind some pratical limitations (availability of track-separated files, ground truth). As far as the fugues keep the strict structure with a clear subject exposition, we are confident that our algorithm should give good results. As an example, the website `http://www.lifl.fr/~giraud/fugues` shows the output on the fugue in Mozart's *Adagio and Fugue in C minor*, K 546. We plan to further experiment it on other baroque or classical fugues, or on more recent corpus such as the Shostakovitch preludes and fugues (op. 87).

References

1. Lisa Browles. Creating a tool to analyse contrapuntal music. Bachelor Dissertation, University of Bristol, 2005.
2. Siglind Bruhn. *J. S. Bach's Well-Tempered Clavier. In-depth Analysis and Interpretation.* 1993. ISBN 962-580-017-4, 962-580-018-2, 962-580-019-0, 962-580-020-4. Available online at `http://www-personal.umich.edu/~siglind/text.htm`.
3. Siglind Bruhn. *J. S. Bachs Wohltemperiertes Klavier, Analyse und Gestaltung.* Edition Gorz, 2006. ISBN 3-938095-05-9.
4. Emilios Cambouropoulos. A general pitch interval representation: Theory and applications. *Journal of New Music Research*, 25(3):231–251, 1996.
5. Emilios Cambouropoulos, Maxime Crochemore, Costas S. Iliopoulos, Manal Mohamed, and Marie-France Sagot. A pattern extraction algorithm for abstract melodic representations that allow partial overlapping of intervallic categories. In *Int. Society for Music Information Retrieval Conf. (ISMIR 2005)*, pages 167–174, 2005.
6. Emilios Cambouropoulos and Costas Tsougras. Influence of musical similarity on melodic segmentation: Representations and algorithms. In *Sound and Music Computing (SMC 04)*, 2004.
7. Raphaël Clifford and Costas S. Iliopoulos. Approximate string matching for music analysis. *Soft. Comput.*, 8(9):597–603, 2004.
8. T. Crawford, C. Iliopoulos, and R. Raman. String matching techniques for musical similarity and melodic recognition. *Computing in Musicology*, 11:71–100, 1998.
9. Asif Ghias, Jonathan Logan, David Chamberlin, and Brian C. Smith. Query by humming: musical information retrieval in an audio database. In *ACM Multimedia*, pages 231–236, 1995.
10. J. Hakenberg. The Pirate Fugues. `http://www.hakenberg.de/music/music.htm`.
11. Walter B. Hewlett. A base-40 number-line representation of musical pitch notation. *Musikometrika*, 4(1-14), 1992.
12. Yuzuru Hiraga. Structural recognition of music by pattern matching. In *International Computer Music Conference (ICMC 97)*, 1997.
13. J. L. Hsu, C. C. Liu, and A. Chen. Efficient repeating pattern finding in music databases. In *International Conference on Information and Knowledge Management (CIKM 1998)*, 1998.
14. Ioannis Karydis, Alexandros Nanopoulos, and Yannis Manolopoulos. Finding maximum-length repeating patterns in music databases. *Multimedia Tools Appl.*, 32:49–71, 2007.

15. Kjell Lemström and Pauli Laine. Musical information retrieval using musical parameters. In *International Computer Music Conference (ICMC '98)*, pages 341–348, 1998.
16. Chih-Chin Liu, Jia-Lien Hsu, and Arbee L.P. Chen. Efficient theme and non-trivial repeating pattern discovering in music databases. In *15th International Conference on Data Engineering (ICDE 99)*, pages 14–21, 1999.
17. Yu lung Lo and Chun yu Chen. Fault tolerant non-trivial repeating pattern discovering for music data. In *IEEE/ACIS International Workshop on Component-Based Software Engineering, Software Architecture and Reuse (ICIS-COMSAR 2006)*, pages 130–135, 2006.
18. Colin Meek and William P Birmingham. Automatic thematic extractor. *Journal of Intelligent Information Systems*, 21(1):9–33, 2003.
19. David Meredith. Pitch spelling algorithms. In *5th Triennal ESOM Conference*, 2003.
20. Marcel Mongeau and David Sankoff. Comparaison of musical sequences. *Computer and the Humanities*, 24:161–175, 1990.
21. Keith S. Orpen and David Huron. Measurement of similarity in music: A quantitative approach for non-parametric representations. *Computers in Music Research*, 4:1–44, 1992.
22. Sami Perttu. Combinatorial pattern matching in musical sequences. Master Thesis, University of Helsinki, 2000.
23. Ebenezer Prout. *Analysis of J.S. Bach's forty-eight fugues (Das Wohltemperirte Clavier)*. E. Ashdown, London, 1910.
24. Dimitrios Rafailidis, Alexandros Nanopoulos, Yannis Manolopoulos, and Emilios Cambouropoulos. Detection of stream segments in symbolic musical data. In *Int. Society for Music Information Retrieval Conf. (ISMIR 2008)*, 2008.
25. Lloyd Smith and Richard Medina. Discovering themes by exact pattern matching. In *Int. Symposium for Music Information Retrieval (ISMIR 2001)*, pages 31–32, 2001.
26. Pei-Hsuan Weng and Arbee L. P. Chen. Automatic musical form analysis. In *International Conference on Digital Archive Technologies (ICDAT 2005)*, 2005.

#	BWV	tonality	voices	S				CS				remarks
				s	g	s'	occ.	cs	g	cs'	occ.	
1	846	C major	4	14	−2	14	21/23					
2	847	C minor	3	20	−2	20	8/8	40	0	40	5/6	
3	848	C# major	3	17	−5	17	12/12	44	0	**42**	7/11	wrong CS
4	849	C# minor	5	5	0	5	14/29	19	+4	19	2/2	
5	850	D major	4	13	−2	**9**	**35**/11	19	0	**15**	8/9	wrong S, S: 24 FP wrong CS, CS: 4 FP
6	851	D minor	3	12	0	12	11/11 $3^i/5^i$	29	0	**33**	2/3	wrong CS
7	852	Eb major	3	16	−8	16	9/9	40	0		0/6	no CS found
8	853	D# minor	3	13	0	**9**	18/19 $7^i/7^i$ (+ 2^i) $3^a/3^a$					wrong S
9	854	E major	3	6	−6	**18**	10/12	22	0		0/⋆	wrong S
10	855	E minor	2	26	+1	26	8/8	36	−2		0/7	no CS found
11	856	F major	3	15	−4	15	10/14	34	0	34	3/5	CS: 1 FP
12	857	F minor	4	11	−3	**10**	10/10	37	0	37	4/8	wrong S, good CS end
13	858	F# major	3	16	−1	16	7/8	41	0	41	2/4	
14	859	F# minor	4	18	0	18	6/7 $2^i/2^i$	44	0	**38**	5/6	wrong CS
15	860	G major	3	31	0	31	4/10 $2^i/3^i$	65	0		0/1	no CS found
16	861	G minor	4	11	0	11	14/16	22	0	22	3/10	
17	862	Ab major	4	7	0	7	15/15			**23**	3/0	wrong CS, CS: 3 FP
18	863	G# minor	4	15	−2	15	12/12	30	0	30	5/7	
19	864	A major	3	13	+6	**11**	**12**/8	21	0		0/2	wrong S, S: 5 FP
20	865	A minor	4	31	0	31	14/14 $5^i/14^i$	44	−12		0/3	no CS found
21	866	Bb major	3	38	0	38	8/8	65	0	65	7/7	
22	867	Bb minor	5	6	0	**10**	11/21			**16**	2/0	wrong S wrong CS, CS: 2 FP
23	868	B major	4	14	0	**13**	10/10 $2^i/2^i$	34	+1	**31**	3/4	wrong S, wrong CS
24	869	B minor	4	21	0	**19**	11/13	45	0		0/3	wrong S, no CS found
							288/306 (29 FP) (85% occ.) $23^i/33^i$ $3^a/3^a$				61/104 (10 FP) (49% occ.)	

^{i} : inverted subject – ^{a} : augmented subject

#7: The correct end for the CS is detected, but the actual CS begins after a small codetta.

#8: The correct end for the CS is detected, but the actual CS begins after a small codetta. Moreover, two incomplete inverted subject (also noted [2]) are detected on measure 54.

#9: The values for CS (⋆) are not counted in the total, as the CS is presented in a segmented form in almost all measures of the fugue [**?**].

Table 2. Results of the proposed algorithm on the 24 fugues of the first book of Bach's *Well-Tempered Clavier*. We take as a truth the analysis of [2], keeping here only the complete occurrence of each pattern. The columns "occ" lists the number of occurrences of Subjects and Counter-Subjects. The values s and cs indicate the index of the note ending the true subject and the counter-subject, whereas s' and cs' are the values predicted by the algorithm. See Section 3 for a definition of g. All false positives (FP) are counted in the remarks.

Enabling Participants to Play Rhythmic Solos Within a Group via Auctions

Arjun Chandra[1], Kristian Nymoen[1], Arve Voldsund[1,2], Alexander Refsum Jensenius[2], Kyrre Glette[1], and Jim Torresen[1]

[1] fourMs, Department of Informatics, University of Oslo, Norway
[2] fourMs, Department of Musicology, University of Oslo, Norway
chandra|krisny|kyrrehg|jimtoer@ifi.uio.no
arve.voldsund|a.r.jensenius@imv.uio.no

Abstract. The paper presents the interactive music system SoloJam, which allows a group of participants with little or no musical training to effectively play together in a "band-like" setting. It allows the participants to take turns playing solos made up of rhythmic pattern sequences. We specify the issue at hand for allowing such participation as being the requirement of *decentralised coherent circulation* of playing solos. This is to be realised by some form of intelligence within the devices used for participation. Here we take inspiration from the Economic Sciences, and propose this intelligence to take the form of making devices possessing the capability of evaluating their *utility* of playing the next solo, the capability of holding *auctions*, and of *bidding* within them. We show that holding auctions and bidding within them enables decentralisation of co-ordinating solo circulation, and a properly designed utility function enables coherence in the musical output. The approach helps achieve decentralised coherent circulation with artificial agents simulating human participants. The effectiveness of the approach is further supported when human users participate. As a result, the approach is shown to be effective at enabling participants with little or no musical training to play together in SoloJam.

Keywords: active music, collaborative performance, conflict resolution, algorithmic auctions

1 Introduction

In many musical cultures and genres there is often a large gap between those who *perform* and those who *perceive* music. In such ecosystems, the performers (musicians) *create* the music, while the perceivers (audience) *receive* the music [11]. Even though perceivers may have some control of the music creation in a concert situation, by means of cheering, shouting, etc., this only indirectly changes the musical output. The divide between performer and perceiver is even larger in the context of recorded music, which is typically mediated through some kind of playback device (CD, MP3 file, etc.). Here the perceiver has very

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

limited possibilities in controlling the musical content besides starting/stopping the playback and adjusting the volume of the musical sound.

The last decades have seen a growing interest in trying to bridge the gap between the performance and the perception of music [6]. Examples of this can be seen as interactive art/museum installations, music games (e.g. Guitar Hero) [7], keyboards with built-in accompaniment functionality [1], "band-in-a-box" types of software, mash-up initiatives of popular artists [10], sonic interaction designs in everyday devices [9], mobile music instruments [2], active listening devices [4, 8], etc. An aim of all such *active music* systems is to give the end user control of the sonic/musical output to a greater or lesser extent, and to allow people with little or no training in traditional musicianship or composition to experience the sensation of "playing" music themselves [5].

There are numerous challenges involved in creating such active music experiences: everything from low-level microsonic control (timbre, texture), mid-level organisation (tones, phrases, melodies) to large-scale compositional strategies (form). In addition comes all the challenges related to how one or more participants can control all of these sonic/musical possibilities through mappings from various types of human input devices. In this paper we will mainly focus on creating a system that is flexible enough for the participants' interaction, yet bound by an underlying compositional idea.

Our approach in SoloJam is to allow for a group of participants with little or no musical training to come together and behave as a "band" of musicians, wherein, they play their respective solos in turn. Thus, the responsibility of playing solos circulates around the band and continues to do so until an indefinite period. To solve the problem of co-ordinating the circulation of responsibility of playing these solos autonomously and effectively, we propose an approach inspired by the Economic Sciences. Specifically, we borrow the concepts of *auctions* and *utility* to address the problem. Our investigation shows that auctions do indeed help decentralised, thus autonomous, circulation of solos within the group. In addition, a careful consideration of the utility function helps participants produce coherent musical output.

We start by introducing the musical scenario that we refer to as SoloJam in Section 2, specifying the issue with participating within it. We then describe our proposed Economics inspired approach to tackling the issue, and the implementation details for the same, in Section 3. Section 4 then looks at the application of the approach within SoloJam, investigating the approach for its effectiveness in enabling participation by artificial agents (who simulate participants with little or no musical training) and human users.

2 The Musical Scenario

In our current context we are interested in creating a system that allows for a group of participants with little or no musical training to get the feeling of being involved with creating music, yet defined in such a way that a certain level of musicality is ensured in the final sounding result. The participants are

to play music using a device that assists them for the same. Such a device, together with the participant using it, is what we call a *node* in this paper. The participant may either be a human user using the device, or an artificial agent behaving in a specified manner simulating a user, and as such associated with the device. In a situation like this, we will need the devices to help co-ordinate the participants' intentions. The devices will have to resolve the conflict that arises from multiple participants wanting to control the same features of the composition. Thus, conflict resolution should be a necessary constituent part of any composition, but indeed, not necessarily the only thing.

In this paper, we focus our attention on this conflict resolution aspect of compositions. As such, we imagine a band of musicians who want to play their respective solos pertaining to the same musical feature. Only one musician ever plays their respective solo at a time. We call this musician the *leader*. However, over time, the playing of solos circulates across the band, as and when other musicians get the opportunity to play their respective solos or become the leader. The control of this circulation of solos happens in a decentralised manner.

The musical space within the system considered in this paper is made up of rhythmic patterns. A sequence of rhythmic patterns when played by one node, is viewed as a *solo* in the context of this paper, until another node commences playing rhythmic patterns. Each rhythmic pattern has a specified number of beats, which we consider as one bar. Thus the musical output is supposed to be a series of rhythmic patterns, one in each bar. Each bar in a sequence can either be a repetition of the rhythmic pattern in the previous bar or not, specifically when played by one node as a solo. And, the next solo, which would be played by another node, should start with a rhythmic pattern that is not exactly the same as, and ideally only slightly different to, the one played by the previous solo playing node in the previous bar. The composition is specified by the aforementioned elements, which also describe the boundaries or constraints to which the musical output should adhere to.

As such, SoloJam can be seen as a compositional idea, or *musical scenario*, where a group of nodes acting in a decentralised fashion come together and take turns in playing a piece based on rhythmic solos. Though nodes act in a decentralised fashion, they must also be able to produce a coherent musical result.

2.1 The Issue With Participation

Given the scenario mentioned above, if a group of participants are to play music, the devices that they use for this participation cannot be traditional instruments. Instead, the devices need to possess some form of artificial intelligence which might allow the group to produce a coherent musical output, and help the participants do so via local interaction with other participants, i.e. without requiring an expert to direct their interaction. Devices helping with coherence are required due to the assumption that the participants do not possess sufficient musical knowledge to produce a satisfactory result on their own. As such, what gets played should be influenced by the devices to some extent, whilst making

sure that the participants are still able to explore the musical space themselves. Devices helping with local interaction are required in order to adhere to the vision of a "band" where members organise themselves into taking turns playing solos, without a central authority directing them. Thus, deciding who plays the solo next should be dealt with by the devices interacting intelligently with each other on behalf of the participants. Such intelligence in the devices forms the crux of the issue with making participants play together effectively within our musical scenario.

We define *decentralised coherent circulation* as giving us a yardstick against which to evaluate the effectiveness of the solution to the issue of allowing participants to effectively play together in SoloJam. *Decentralisation* means that there is no central control over the circulation of playing of solos by participants. *Coherence* in our case means for nodes to be playing slight variations of each others rhythmic patterns over time as and when they become leaders, such that the next leader plays a slight variation of the rhythmic pattern played by the current leader. Thus, our goal is to design an intelligent system that allows for both decentralised control and coherent musical output. It should enable participants to play together without them possessing much musical knowledge, and without requiring an expert to direct their interaction with other participants.

3 Economics Inspired Approach For Participation

We now propose the approach inspired by the Economic Sciences to tackle the issue described in Section 2.1. A detailed specification of this approach follows.

3.1 Specification of the Approach

One can see the problem of decentralised control of circulation of solos as a resource allocation problem, where the resource can be viewed as a metaphor for *having the responsibility of playing a solo*. This responsibility is what needs to be continuously allocated to the node who may be *most deserving* of being the leader within SoloJam at any point in time.

The concept of auctions has a long standing history in human society, where the idea is to have a mechanism in place that allows for the allocation of resources/goods/services via the exchange of these resources/goods/services with other resources/goods/services, or indeed some currency. Anything that may be exchanged has some value for the parties between which the exchange happens. This is where the concept of utility comes in. Utility [3, 12], as a concept, has a long history in the Economic Sciences as being an idea that allows for expressing the value of a choice or decision that one needs to make, for example, how much may one be willing to spend in choosing to buy a guitar is the value of the guitar for the individual. This value can, with certain assumptions about the preferences of the individual with respect to making choices, be quantified in the form of a mathematical function. Such numerical expression of value makes exchanging resources/goods/services practical.

Assuming that it may be possible to compute the deservedness of being the leader, at every time step, whilst the leader is playing its solo, we make it also hold (broadcast) an *auction*, in which all other nodes can *bid* in order to become the next leader. We thus design the node such that every node can evaluate the deservedness of it being the leader as a *utility* of its current rhythmic pattern. The utility values of their respective rhythmic patterns are what the nodes use as their respective bids. As such, at any given time, the node with the highest utility for their respective rhythmic pattern, must be the leader, provided this value is computed truthfully (or honestly). At every time step, the bidder nodes can also change their respective rhythmic pattern, in order to come up with a new rhythmic pattern with higher utility as compared to the utility of their current rhythmic pattern. The transfer of responsibility happens when a bidder node wins the auction held by the leader. This necessitates a gain for the leader, i.e. the auction can only be won if the leader gains from handing over the responsibility to the highest bidder. This implies that the utility of the rhythmic pattern that the leader is currently playing, must, at the time of the transfer, be lower than the highest bid it receives. We now detail the auction mechanism for decentralised circulation of responsibility, and the concept of utility for computing deservedness and coherence.

Auction Mechanism. The leader holds a second-price sealed-bid auction, in particular, the *Vickrey auction* [13] in every bar, to receive bids from the bidders which then are used to decide whether or not there is a winner to whom the responsibility of playing the solo would pass in the next bar. The reason for this design choice is that Vickrey auctions deem truthful bidding to be the dominate bidding strategy. In our case, this means that a bidder can do no better than bidding with the true utility value of their rhythmic pattern. The second-price nature of the auction suggests for the winner of the auction to make a payment equal to the value of the second highest bid to the leader. This second price aspect of this auction mechanism makes truthful bidding a dominate bidding strategy. However, in the current setup we do not exchange money[3] (in the form of such payments by bidders to the leader). This means that, although the transfer of responsibility necessitates a gain for the leader, as mentioned above, the leader only ever compares the received bids and the current utility of its own rhythmic pattern, in order to ascertain whether or not it should hand over the responsibility to the highest bidder. Ties in bids, when the bids are higher than the leader's rhythmic pattern utility, are broken randomly. The sealed-bid nature of the auction requires that the bids are not public and only known to the bidder and leader. We leave the consideration of exchange of money and other possibilities offered by this auction mechanism to the future, when dealing with more complex variants of SoloJam.

[3] The auction and bidding setup in SoloJam allow for money (or virtual money), in the form of bid values to be exchanged. But, we only consider monitoring the utilities for now.

Utility. To participate in the auction effectively, each node must have a way to evaluate and communicate a value that it considers its current rhythmic pattern to be worth. A rhythmic pattern in SoloJam is represented as a bit string parsed from left to right, whereby, a 1 indicates 'triggering a beat' and a 0 represents 'not triggering a beat'. For each node, we define a utility function which the node uses to evaluate the value of its current rhythmic pattern, both in relation to itself and to the leader, knowing its role as either a bidder or a leader. The following equation specifies part of this utility function:

$$u_i = \frac{c}{(1 + aD_l)(1 + bT_l)} \tag{1}$$

Here, D_l is the hamming distance of a node's current rhythmic pattern with respect to the leader's current rhythmic pattern, T_l is the length of time a node has been playing the solo, i.e. the number of bars a node has played rhythmic patterns as a leader, the coefficient a is the importance (in terms of a weighting) given to D_l, the coefficient b is the importance (in terms of a weighting) given to T_l, and c is a normalisation constant. In addition to this, two more conditions completely specify the utility function. These clauses being:

1. The utility is *zero* for a bidder node if D_l goes below $\epsilon\lambda$, where ϵ is a small percentage of the length of the rhythmic pattern (λ).
2. The utility is *zero* for a bidder node if the node has handed over control to a new leader node in the previous time step.

According to the utility function above, the longer (in terms of bars) a node plays a solo as the leader, the lesser it values its current rhythm, indicating boredom or fatigue, of which the node is made aware via the utility function. The node also possesses knowledge about the hamming distance between its own and the leader's respective rhythmic patterns. This knowledge can be used by the node to come up with rhythmic patterns that are of higher value, given the leader's rhythmic pattern. The closer a node can match its rhythmic pattern against the leader's pattern, the higher the node values its own pattern. This remains true as long as the match does not get closer than or equal to $\epsilon\lambda$, allowing for the node to stir clear of intending to play a rhythmic pattern that may be very similar to or exactly the same as that of the leader (as per the first clause above). Additionally, we can see that this specification of utility, taking the leader's rhythmic pattern into consideration, also provides the node with a gradient (i.e. the closer the rhythmic pattern to that of the leader, the higher its value), which it may make available to the participant in order for them to come up with rhythmic patterns which are slight variations (at least $\epsilon\lambda$ different) of the leader's rhythmic pattern. As such, in addition to computing deservedness, we see the utility function as a means of instilling coherence in the musical output from SoloJam. Note that D_l forms the main link between nodes (the node in question and the current leader node), and the coefficient a associated with D_l emphasises or otherwise, the strength of this link. We will put this coefficient to use for the investigation carried out in this paper in Section

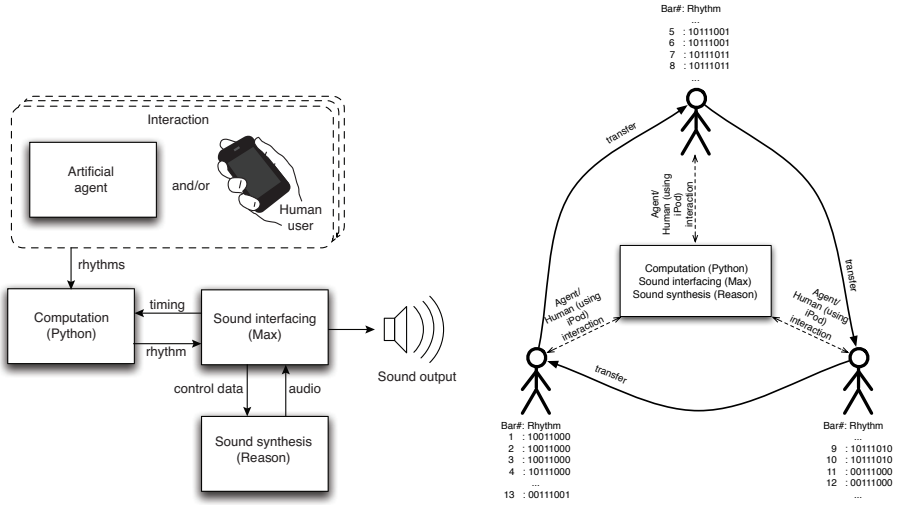
4. The clauses above further indicate a way to carefully consider designing the utility function in order for a globally coherent piece of music to result from local interactions within SoloJam. The first clause suggests for there not to be a perpetual repetition of the same rhythmic pattern by all the nodes of SoloJam, which would be monotonous. The second clause allows for a node to not take over the responsibility soon after it released it, which may happen otherwise, since the node's rhythmic pattern would already be a slight variation of the new leader that took over the responsibility from this node. Not considering this clause may thus reduce the variations that may occur in the music performance in the global sense.

3.2 Implementation

Fig. 1 shows the building blocks of the implementation of SoloJam. Fig. 1(a) outlines the schematic of the implementation of SoloJam. The current SoloJam scenario has been implemented on a Macintosh computer, in conjunction with iOS devices for human interaction within the scenario. The setup can be broken down into 4 modules: the Computation module, the Interaction module, the Sound interfacing module, and the Sound synthesis module.

The Computation module is implemented in Python and simulates our approach for effective participation described in Section 3.1, with a thread representing each node. These threads interface with the Interaction module as well as the Sound interfacing module. The Interaction module can function in two ways. If an artificial agent is to be part of the node, the thread in the Computation module representing this node is made to implement the functionality of the agent in terms of the manner in which this agent comes up with rhythmic patterns. If a human user is to be part of the nodes, iOS devices (specifically iPod Touch) are used for sensing human motion, and specifically for SoloJam, sensing the shaking of the device (using the built-in inertial sensors). The signals from shaking are sent as Open Sound Control (OSC) [14] messages to a thread in the Computation module associated with the device, which are then converted into rhythmic patterns within this thread. The bit strings representing rhythmic patterns are further sent as OSC messages to the Sound interfacing module, together with the utilities/bids (computed within the Computation module) of leader/bidder node rhythmic patterns at every bar.

The Sound interfacing module is implemented as a Max/MSP patch. It serves as a control module for the SoloJam scenario, accepting strings of rhythmic patterns, synchronising and converting them to control signals for the Sound synthesis module. The audio streams from the Sound synthesis module are channeled back to the Sound interfacing module for mixing and effects processing. The Sound interfacing module also performs a visualisation of various aspects of the system, such as node utilities. The Sound synthesis module is currently instantiated as a virtual sound module rack in Reason. A drum kit synthesiser module is used for each node. Reason is controlled by the Sound interfacing module through ReWire. MIDI signals are sent to the synthesisers, and the audio streams are sent back to the Sound interfacing module.



(a) Implementation. (b) Illustration.

Fig. 1. Building blocks of the implementation of SoloJam showing (a) a schematic of the implementation of SoloJam, and (b) an illustration of the SoloJam scenario within the context of this implementation.

Fig. 1(b) illustrates the SoloJam scenario within the context of the aforementioned implementation. It shows 3 agents or human users participating in the scenario. The rhythmic patterns associated with each participant at various bars are shown. These rhythmic patterns are fed in to our auction based approach for effective participation simulated by the Computation module. As per the rhythmic patterns shown, one possibility for the transfers of responsibility of playing solos is indicated in the figure.

4 SoloJam with Participants

We now look at how the auction based approach proposed in this paper, together with the proper design of the utility function, enables effective participation within the composition. We primarily look at the case where artificial agents are considered as simulating the behaviour of participants with little or no musical training, and act within SoloJam as participants. The case where SoloJam involves human participants is also discussed.

4.1 SoloJam with Artificial Agents: Enabling Participation

Although SoloJam involves human interaction, in order for behavioural equivalence across the participants, we consider experimenting with artificial agents in this section. Moreover, an artificial agent can be designed to behave as a participant with little or no musical training with little effort. As such, we get

artificial participants behaving in a specified manner operating the respective nodes similarly. This allows for evaluating a base line system, which is a system that should work even if all the nodes are operated by participants with little or no musical training. Otherwise, one could argue that a human operator may influence the system towards having the requisite functionality, even if the system did not work. Thus, artificial agents allow for controlling the nature of the interaction of the operator, removing human induced functionality into the circulation of solos, which may be hard to account for.

We primarily investigated the effects of the utility function specification within SoloJam, considering the manner in which knowledge about the leader node affects the circulation of solos within the group of participating nodes. Since we are only interested in the effect of the utility function on the circulation, fixing other factors which may influence the circulation, makes a plausible case for using artificial agents with a fixed behaviour. In this study, these artificial agents use the notion of mutation to generate the bit strings that represent rhythmic patterns. This mutation is such that the agents can flip each bit in their bit string with a probability $1/\lambda$, where λ is the length of the rhythmic pattern. In so doing, the agent generates a new rhythmic pattern, which is a mutation of its old rhythmic pattern. This mutation based rhythmic pattern generation process is essentially used by bidder nodes in every bar they have to bid in, as they search for slight variations of the leader's rhythmic pattern. We limit our study with agents to the case where, once the leader starts playing their solo, they do not change their rhythmic pattern for the duration of the solo (which should be some bars long), i.e. a solo is made up of repetitions of the same rhythmic pattern. This limitation allows us to clearly observe if the bidder nodes are indeed able to search for slight variations of the leader's rhythmic pattern, which, upon winning the auction, they eventually play.

Note that the coefficient a, within Equation 1, signifies the importance (in terms of a weighting) that a node gives to the distance D_l between its current rhythmic pattern and the leader's current pattern. Setting the value of this coefficient to 0.0 within a node, allows for switching off knowledge about the leader node. In essence, the node then only knows its own rhythmic pattern and the duration it has played a rhythmic pattern when acting as a leader. Setting a to a positive value makes the node consider knowledge about the leader. We take $a = 0.0$ and $a = 1.0$ in order to explicitly investigate the effects of not disclosing and disclosing respectively, the knowledge about the leader node to other nodes. Note that the leader node remains unaffected from a change in the value of a, because D_l is zero for it, thus making a irrelevant.

We can now detail the effects of such knowledge within the workings of SoloJam, specifically looking at the nature of the decentralised circulation of solos and also the coherence that can be achieved in the generated piece of music. We first look at the piece resulting from the system, and then provide a discussion based on the evolution of the utilities of the nodes, both with respect to such knowledge. For our study, we use the following parameter settings: *Rhythmic pattern length* (λ) $= 8$, $\epsilon = 0.1$, *Node count* $= 3$, $c = 2$, $b = 0.05$.

Observations About the Resultant Piece. Figs. 2 and 3 show snapshots of rhythmic patterns that are generated when the agents play SoloJam, under two specific cases, one where bidder nodes do not consider using knowledge about the leader's rhythmic pattern when evaluating their own rhythmic patterns, and the other where they do so. These two cases are realised by $a = 0.0$ (Fig. 2) and $a = 1.0$ (Fig. 3) respectively within the part of the utility function (Equation 1) used by each node for this evaluation.

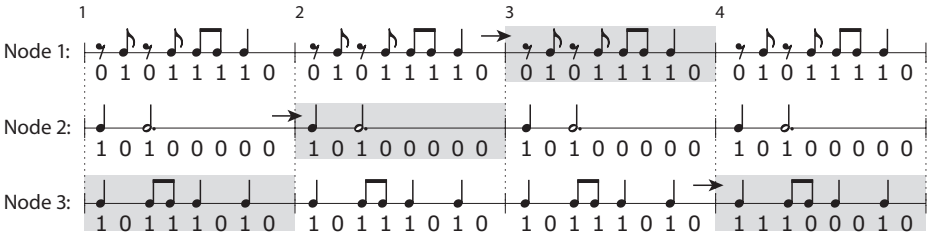


Fig. 2. Snapshot of the rhythmic patterns when $a = 0.0$. There is maximal circulation of responsibility of playing solos (at every bar). The musical output is incoherent as there is no mutation towards closer rhythmic patterns by bidders. In effect, there is no active participation via mutation. Enabling participation is not effective.

These figures show the rhythmic patterns as bit strings and in music notation, for each node in the system. Since we have 3 nodes, 3 lines with bit strings and music notation correspond to each node, as indicated. These lines can be read from left to right for each node. At the end of the 3 lines, the reader can continue at the left of the next 3 lines (see Fig. 3), and so on. Each bar is clearly marked as enclosing the respective rhythmic patterns (of length 8 bits) for each node. The shaded regions denote the current leader. An arrow between bars denotes a rhythmic pattern being sufficient for a transfer to happen. Mutations within a pattern from a previous bar for a node are denoted by dotted circles. The numbers above bars are bar numbers, wherein a range means that the rhythmic pattern is repeated for all the bars in that range, without any mutations or transfers.

For the case with $a = 0.0$, the 3 nodes do not mutate their respective rhythmic patterns over successive bars. Moreover, the transfer of control of responsibility for the solo happens in every bar, as indicated by the shaded regions in the figure. For the case with $a = 1.0$, we can see a more interesting final result: it can be seen that at bar 21, the rhythmic pattern with which Node 1 bids in the auction held by Node 3 (the then leader), varies less (different by 1 bit) from Node 3's rhythmic pattern, as compared to the rhythmic pattern associated with Node 2 (different by 2 bits). Node 1 wins this auction in this bar, and from bar 22 onwards until bar 42, plays its rhythmic pattern. At bar 23, Node 2 and 3 mutate their rhythmic patterns, a further mutation happening at bar 29 for Node 3. Note that in bar 23, Node 3 comes up with a rhythmic pattern that is

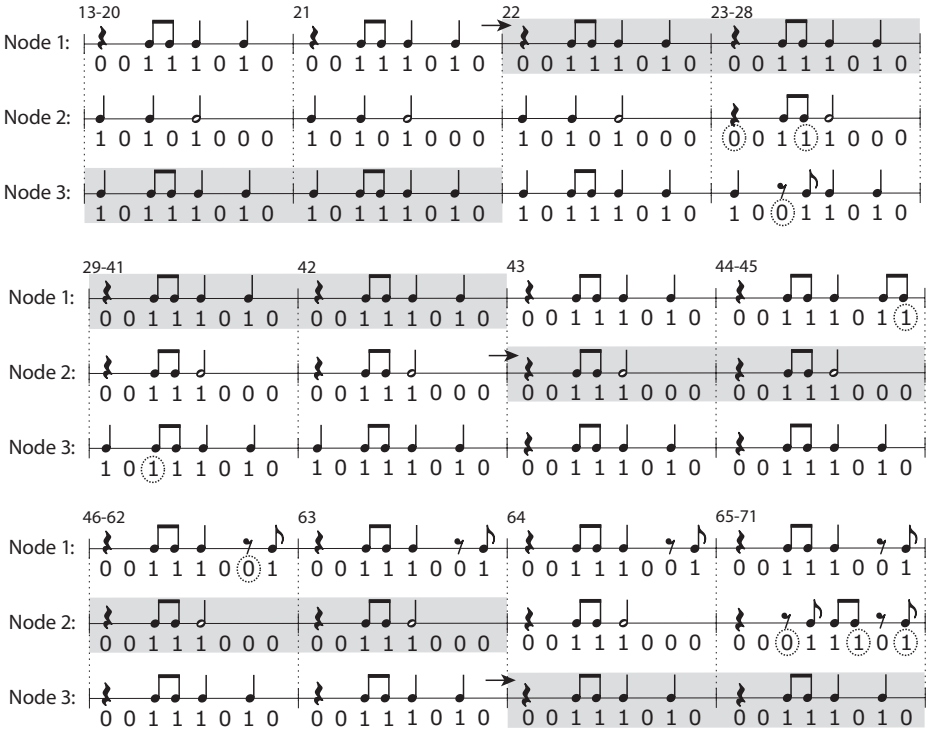
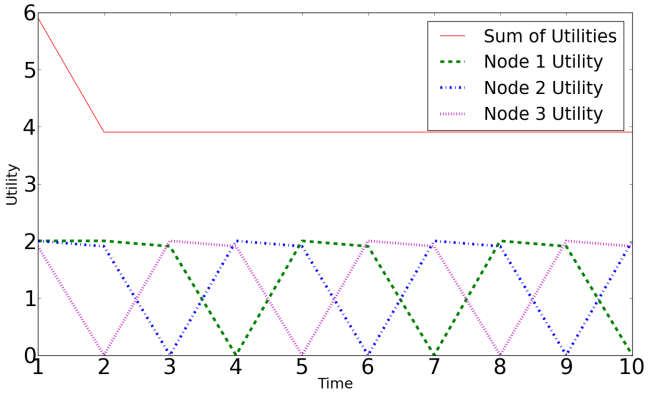


Fig. 3. Snapshot of the rhythmic patterns when $a = 1.0$. Decentralised coherent control is exhibited. The circulation of responsibility of playing solos happens after the leader having played their rhythmic pattern for some bars. Coherence results from the nodes actively searching for closer variants, via mutation, of the leader's rhythmic pattern, and the closest rhythmic pattern being played by the respective bidder, provided the bidder wins the auction. Enabling participation is effective.

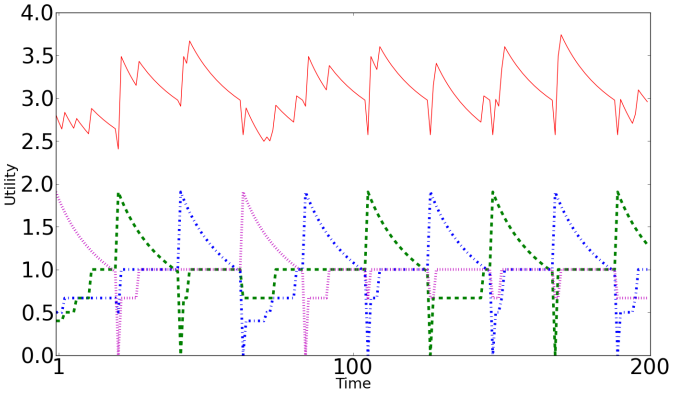
2 bits different from the leader, as compared to its rhythmic pattern in bar 22. This is because the rhythmic pattern in bar 22 has its value reduced to zero in the following bar in accordance with the utility function. Thus, any mutation of that rhythmic pattern in the bar following that will have a value greater than zero. As such, this mutation will replace the previous rhythmic pattern. Other than such a situation, the mutations that are generated over time take the nodes closer to the rhythmic pattern of the leader, as can be seen in the figure. In bar 42, there is a tie between Node 2 and Node 3, which is broken randomly and Node 2 takes over the responsibility of playing its rhythmic pattern as a solo. In bar 44, and then in 46, Node 1 mutates towards a closer variant of Node 2. This is followed by a tie again in bar 63, which is then broken randomly in favour of Node 3. In bar 65, Node 2 mutates away from Node 3, again due to the nature of our utility function, as described above. It is clear from Fig. 3 that the nodes actively search for closer variations of the leader's rhythmic pattern via

mutation, and the node (that is sometimes decided upon by a tie break) with the closest match, becomes the leader in the next bar, provided this node wins the auction.

Discussion Based on the Utilities of Nodes. Fig. 4 plots the utilities of the rhythmic patterns of each node, as individually evaluated by these nodes using the utility function described in Section 3.1, and the sum of these utilities. These figures correspond to the snapshots of the pieces from the system (Figs. 2 and 3).



(a) $a = 0.0$



(b) $a = 1.0$

Fig. 4. Utilities of nodes (a) without ($a = 0.0$) and (b) with ($a = 1.0$) knowledge about the leader's rhythmic pattern.

As observed with the corresponding piece (Fig. 2), for the case when $a = 0.0$, the transfer of control happens at every time step, thus a leader node only ever plays its rhythmic pattern for one bar. The auction held by the leader

immediately leads to the bidder who was not the previous leader, to take over the control from the leader, thus becoming the new leader, but for only one bar. This happens due to the nodes not considering using the knowledge about the leader's rhythmic pattern, and thus having a utility and bid of $u = c = 2.0$, if they were not the leader in the immediate previous time step. The process of such transfer of controls carries on. Note that all possible mutations of rhythmic patterns for a bidder who was not the leader in the previous bar, have the same value of 2.0. Thus, the agent has no pressure towards coming up with bids of higher value. We see however, that there is not enough time for the bidders to search (via mutation) for new rhythmic patterns. This is because when a mutation results in a new rhythmic pattern, the previous rhythmic pattern has its value equal to the value of this new rhythmic pattern at all times, be it in the round after the round in which the node was the leader (the value for both rhythmic patterns is 0.0 in this case), or the rounds after this (value is 2.0). As such, the rhythmic patterns with which the nodes started with in the first bar, either as a leader or bidder, remain as the rhythmic patterns associated with these nodes forever, as can also be observed in the corresponding piece for the $a = 0.0$ case (Fig. 2). In effect, coherence remains an issue, since the initial rhythmic patterns of the nodes will not necessarily be slight variations of each other. Moreover, the fact that nodes play their rhythmic patterns for only one bar, goes against the whole idea behind playing solos, unless of course playing for only one bar were to be a requirement from the composition. Most importantly, however, the agents are not able to actively participate to explore the composition. The current utility function with $a = 0.0$ is thus not suitable for being used when participants are to play rhythmic solos within a band-like setting, or else there would be maximal circulation of control (at every bar, thus no solo being played), the musical output will be incoherent, and there would be no active participation.

For the case when $a = 1.0$ however, the playing of solos and transfer of control over time happens in a more favourable manner with respect to the envisaged goal of local interaction producing a resultant globally coherent piece of music, or decentralised coherent circulation. Fig. 4(b) shows spikes in the node utilities, which indicate the start of nodes playing their rhythmic pattern as solos, and these utilities deplete over time. Whilst the leader node's rhythmic pattern utility depletes, the bidder nodes have their artificial agents search towards slight variations of the leader's rhythmic pattern, as indicated by the increase in their utilities over time. As a result, the leader gets to play its rhythmic pattern as a solo for some time and then hands over control to the bidder managing to search and bid to play the closest variation of the leader's rhythm, as observed with the corresponding piece in Fig. 3. The flat regions in the utility graphs (Fig. 4) indicate agents associated with bidder nodes having found rhythmic patterns at a distance D_l of $\epsilon\lambda$ from the leader's rhythmic pattern. Note that there are always multiple rhythmic patterns that the agent could come up with, all of which differing by distance $\epsilon\lambda$ from the leader's rhythmic pattern, which can be seen as the flexibility in the composition that may be explored by a participant based

on their preferences, e.g. preferring one rhythmic pattern over another, even though these rhythmic patterns have the same utility assigned to them by the device. The artificial agents mimicking participants have thus been enabled to play rhythmic solos in a decentralised and coherent fashion via the consideration of a utility function that takes the knowledge of the leader's rhythmic pattern into account. The agents must now, as compared to the case where $a = 0.0$, actively participate to search for a rhythmic pattern, and upon being the leader, play them. The solos that get played adhere to the composer defined boundaries as defined in Section 2, and the system maintains a decentralised coherent circulation. As mentioned before, having a decentralised coherent circulation shows that the system enables the agents to play through the composition effectively.

It would be interesting to consider how the increase in the number of nodes affects the resultant behaviour of the system, with nodes possessing a utility function such as the one defined in this paper, for the case with $a = 1.0$. We leave this as future work.

4.2 SoloJam with Human Users

SoloJam with human participation was also implemented. As mentioned before, human participation involves a human user using a device that allows for the exploration of the composition. The iPod Touch devices that we use for human participation, one for each human user, have a thread each in the Computation module representing them. Upon shaking the device, the signals from this shaking are received by the associated thread and converted into a rhythmic pattern, which becomes the candidate rhythmic pattern for the next bar for the node in question. The human user, unlike the agent, may change the rhythmic pattern in any bar when part of a leader node.

A video of SoloJam with human participation can be found online[4]. The first part of the video shows some sounds and sound effects played together with the sound output of SoloJam. These sounds and effects are part of a further extension of the SoloJam scenario, and are not relevant to the discussion of this paper, thus we leave their discussion to the future. The second part shows three people using iPod Touch devices to play through the piece, playing rhythmic solos as leaders, and bidding for playing slight variations of the leader's current rhythmic pattern as bidders. Fig. 5 shows a labelled screenshot of this video. The Max/MSP patch (our Sound interfacing module described in Section 3.2) in the background visualises the utilities (three horizontal bars at the top right part of the patch) for each node. The top horizontal bar is the utility associated with the person on the right (Node 1). The middle horizontal bar is associated with the person on the left at the back (Node 2). The lower horizontal bar is associated with the person on the left at the front (Node 3). The reader is advised to only focus on the rhythmic patterns resulting from the users shaking their devices and the horizontal bars representing utility for each node. The circulation of solos in this particular video follows the sequence: *Node 3* $\rightarrow$ *Node 1* $\rightarrow$ *Node 2*.

[4] `http://www.fourms.uio.no/videos/SoloJam.mov`

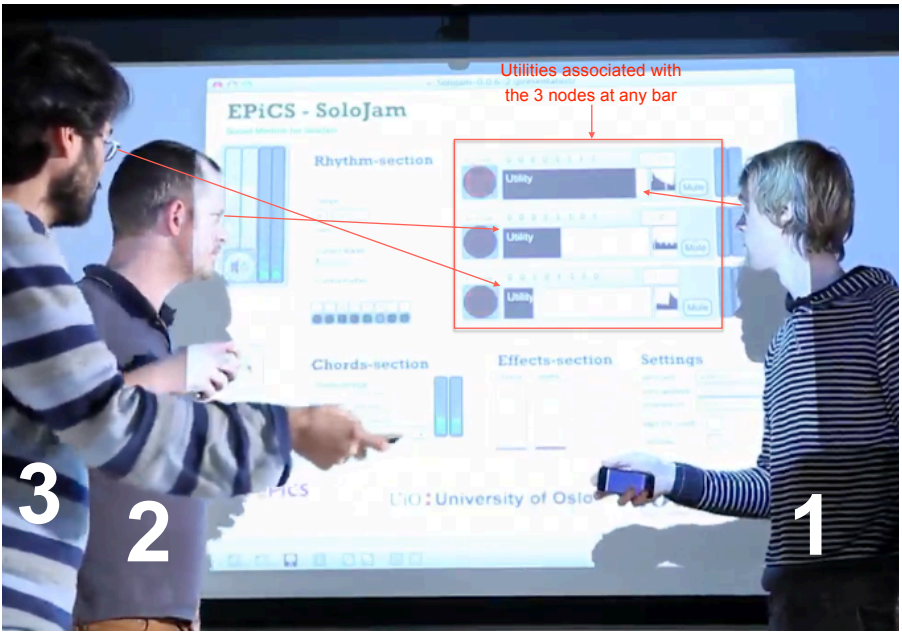


Fig. 5. Labelled screenshot of the video of SoloJam with human participation.

In the video, it is possible to see that the bidder node whose rhythmic pattern is closest to that of the leader node, has a utility higher than that of the other. Also, the transfer of responsibility happens when the leader node's utility goes below the utility of the highest bidder node. Furthermore, the bidder nodes have their rhythmic pattern utilities increased, as and when they come up with closer (in terms of hamming distance) variations (but not exact copies) of the leader's rhythmic pattern. Thus, our approach encourages human users to come up with rhythmic patterns that are slight variations of the leader. The closest bidder is then aptly rewarded by this bidder becoming the leader, once this bidder wins the auction held by the leader.

5 Conclusions

We have outlined and discussed the issue with enabling participants with little or no musical training to play together in the interactive music system SoloJam. An approach inspired by the Economic sciences, specifically borrowing the concepts of auctions and utility, is proposed in order to address this issue. Nodes that possess the capability of evaluating the deservedness of being able to take on the responsibility of playing the solo starting in the next bar (via a utility function), and auctioning and bidding capabilities, are shown to exhibit decentralised co-ordination when circulating solos in SoloJam. Furthermore, a careful

design of the utility function enables participants (simulated by artificial agents) to come up with an output that is musically coherent. This is highlighted by the manner in which the agents, as bidders, search towards higher utility variants of the leader node's rhythmic pattern. These variants, in fact, are slight variations of the leader node's rhythmic pattern. We further exhibit human user participation within SoloJam supporting our approach. In effect, decentralised coherent circulation that results from our Economics inspired approach, demonstrates the effectiveness of the approach towards enabling participation within SoloJam. Having proven the concept, our next step will be to conduct usability tests with human participants having different types of musical backgrounds. In addition to testing the system with participants with little or no musical training, we are also interested in seeing how music students and professional musicians interact with the system.

Acknowledgments. The research leading to these results has received funding from the European Union Seventh Framework Programme under grant agreement n^o 257906 and the Norwegian Research Council through the project Sensing Music-related Actions (project n^o 183180).

References

1. Bunce, G.: Electronic Keyboards: their use and application in Secondary School Music teaching. Master's thesis, Royal Holloway, University of London (2005)
2. Essl, G., Rohs, M.: Interactivity for mobile music-making. Organised Sound 14(2), 197–207 (2009)
3. Fishburn, P.C.: Utility theory. Management Science 14(5), 335–378 (1968)
4. Goto, M.: Active music listening interfaces based on signal processing. In: IEEE International Conference on Acoustics, Speech and Signal Processing. vol. 4, pp. 1441–1444 (2007)
5. Jennings, K.: Toy symphony: An international music technology project for children. Music Education International 2, 3–21 (2003)
6. Leman, M.: Embodied Music Cognition and Mediation Technology. MIT Press, Cambridge, Massachusetts (2008)
7. Miller, K.: Schizophonic performance: Guitar hero, rock band, and virtual virtuosity. Journal of the Society for American Music 3(4), 395–429 (2009)
8. Moens, B., van Noorden, L., Leman, M.: D-jogger: Syncing music with walking. In: Sound and Music Computing Conference. pp. 451–456. Barcelona, Spain (2010)
9. Rocchesso, D.: Explorations in Sonic Interaction Design. Logos, Berlin (2011)
10. Shiga, J.: Copy-and-persist: The logic of mash-up culture. Critical Studies in Media Communication 24(2), 93–114 (2007)
11. Small, C.: Musicking: The Meanings of Performing and Listening. Wesleyan University Press, Hanover, New Hampshire (1998)
12. Stigler, G.J.: The development of utility theory, Parts I and II. Journal of Political Economy 58, 307–327 and 373–396 (1950)
13. Vickrey, W.: Counterspeculation, auctions, and competitive sealed tenders. The Journal of Finance 16(1), 8–37 (1961)
14. Wright, M.: Open Sound Control: an enabling technology for musical networking. Organised Sound 10(3), 193–200 (2005)

Demo session

Development of a Test to Objectively Assess Perceptual Musical Abilities

Lily Law[1] and Marcel Zentner[2],

[1,2] Department of Psychology,
University of York, Heslington,
York, YO10 5DD
UK
{l.law, m.zentner}@psych.york.ac.uk

Abstract. Possibilities for a fine-grained and objective measurement of individual differences in musical abilities are limited at present. A common approach to determining musical competence therefore is to rely on information about the extent of individuals' musical training. Yet relying on musicianship fails to identify musically untrained individuals with musical skill. To counteract this limitation, we developed a test-battery which can be taken by musicians and non-musicians alike, and which measures perceptual musical skills across multiple domains: Tonal (Melody, Pitch), Qualitative (Timbre, Tuning), Temporal (Rhythm, Rhythm-to-Melody, Accent, Tempo) and Dynamic (Loudness).

Keywords: music, ability, talent, perception, individual differences, melody, pitch, timbre, tuning, rhythm, accent, tempo, loudness

1 The Development of a New Music Test

Across disciplines, scholars are increasingly interested in assessing and understanding individual differences in musical ability. One reason for the current interest in music and the mind are the relationships between musical abilities to non-musical traits, ranging from empathy to dyslexia. For example, problems in rhythm perception have been recently found to relate to reading impairments, and there is reason to believe that training of rhythmic processing capacities could act as a remedy for dyslexia [1]. A more complete picture of the links between musical and non-musical traits may also shed light on another hotly debated issue, the evolutionary origins of music. Unfortunately, such a tool does not currently exist. It is not that various aspects of music perception and production had not been extensively investigated – they have [e.g., 2]. What has been missing is interest in the development of a psychometrically sound and construct validated test, capable of diagnosing individual differences in musical ability. Most musical aptitude tests were developed between 1920 and 1970 and originated in music education research. The primary goal of these tests was to identify the potential for musical accomplishment in young children [e.g., 3, 4, 5, 6, 7]. These tests however are inaccessible or have not proven to be useful enough for use in contemporary research.

We have developed a new test-battery using a web-platform, for measuring individual differences in perceptual musical skills, rectifying some of the shortcomings in earlier tests. To this end, we felt it necessary to expand the range of perceptual musical skills usually omitted from previous tests. In addition to tasks testing tonal memory and rhythmic skills, our battery includes tasks testing skills in the perception of tempo, timbre, tuning, pitch, accent, and loudness. These parameters

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

are considered of prime importance in the expression and perception of musical performances [8, 9]. Our test-battery provides a music background questionnaire and a non-music related questionnaire in addition to the listening test. This not only pins down the specific music factor(s) or experience that might have facilitated music perceptual skill, but also further informs us as to whether there are other non-musical activities that share transferable skill properties. Also worth noting are the main differences between our test-battery and the Goldsmiths Musical Sophistication Index (Gold-MSI) [10] – specifically that our test-battery measures nine basic judgments on perceptual listening skill (using the same/different paradigm) whereas the Gold-MSI measures perceptual skills on melody discrimination, rhythm accuracy and musical genre classification modeled on previous research. Although the authors of Gold-MSI have provided evidence of internal consistency reliability for their survey questionnaire, other areas of psychometric properties such as test-retest reliability and test-validity are yet to be reported either for the music test or survey questionnaires. On the other hand, our test-battery has shown to be reliable and validated.

It has showed satisfactory psychometric properties for the composite score (internal consistency and tests-retest analysis >= .89) and fair to good ones for the individual subtests (.62 - .83) (see Table 1). Convergent validity was established with the relevant dimensions of Gordon's Advanced Measures of Music Audiation [11] and Musical Aptitude Profiles (Melody, Rhythm, Tempo) [4], the Musical Ear Test (Rhythm)[12], and content validity with sample instrumental sounds (Timbre) (see Table 2). There was a moderately strong relationship between test performance and self-reported musical training, providing additional support to the test's validity but also suggesting that the current instrument accounts for variance in musical skills beyond self-reported musicianship status and previous musical aptitude tests. The current work also suggests that performance on the nine subtests may be subtended by two higher order perceptual abilities: an analytical and a sensory perceptual musical ability. The analytical factor is related to memory capacity and the sensory factor refers to quick attention capacity and judgment. We also found the rhythm subtests from our test-battery are related to spatial or logic reasoning ability.

This new test-battery is very useful in many ways. First, the battery is more comprehensive compared to previous tests comprising of nine music perceptual tests as well as music background questionnaires. Thus it is a potential tool for investigating a wider range of perceptual skills and can go beyond the conventional focus on rhythm and tonal memory. Second, as the test-battery was developed more recently than previous tests, it is more sensitive with current concepts of music perception and cognition. Third, high standards for test construction and validation were applied.

In conclusion, we hope that the current battery can provide a basis from which a richer scientific narrative on musical ability and its measurement will eventually emerge.

Table 1. Split-Half Reliability and Test-Retest Coefficient for Subtests and Composite Score

Test	Internal Consistency	Test-Retest
Tuning	.82	.68**
Rhythm-to-Melody	.80	.82**
Pitch	.78	.77**
Timbre	.73	.68**
Melody	.71	.77**
Loudness	.68	.83**
Rhythm	.67	.62**
Accent	.66	.71**
Tempo	.64	.81**
COMPOSITE	.89	.90**

Note. Sample size for internal consistency was N=56; sample size for Test-Retest was N=20; ***p*< 0.01 (2 tailed)

Table 2. Correlation between AMMA, MET, MAP, Timbre (Mono) with the New Music Test (Convergent and Content Validity)

Test	Tonal (AMMA)	Rhythm (AMMA)	Rhythm (MET)	Tempo (MAP)	Timbre (Mono)
Melody	**.68****	.60**	.46**	.60**	.23
Rhythm-to-Melody	.43**	**.42****	**.64****	.44**	.33*
Rhythm	.51**	.44**	**.60****	.37**	.23
Accent	.48**	.37**	.37**	.44**	.24
Tempo	.33*	.33*	.22	**.33***	.36**
Timbre	.30*	.27	.15	.32*	**.53****
Tuning	.48**	.41**	.28*	.47**	**.41****
Pitch	.34*	.33*	.12	.37**	**.49****
Loudness	-.10	-.11	-.05	.05	**.40****

Note. N=52; Target validity correlations are in bold fonts. AMMA= Advanced Measures of Music Audiation, MET= Musical Ear Test, MAP = Musical Aptitude Profile
* *p*< 0.05; ***p*< 0.01 (2 tailed)

References

1. Thomson J.M., & Goswami U.: Rhythmic Processing in Children with Developmental Dyslexia: Auditory and Motor Rhythms Link to Reading and Spelling. J. Physiol. - Paris, 102, 120-129 (2008)
2. Jones, M.R., Fay, R.R., & Popper, A.N.: Music Perception. Springer, New York (2010)
3. Bentley, A.: Musical Ability in Children and Its Measurement. Harrap, London (1966)

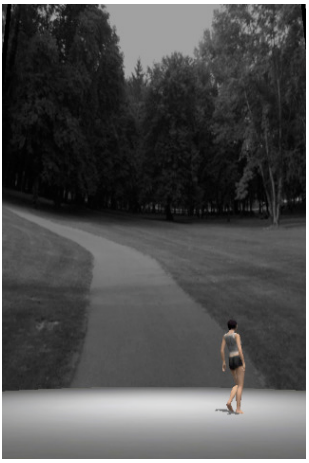
4. Gordon, E. E.: The Musical Aptitude Profile: A New and Unique Musical Aptitude Test Battery. CRME. 6, 12–16 (1965)
5. Karma, K.: The Ability To Structure Acoustic Material As A Measure of Musical Aptitude. 1. Background Theory and Pilot Studies. Research Bulletin No. 52, Institute of Education, University of Helsinki (1973)
6. Wing, H.: Tests of Musical Ability and Appreciation: An Investigation Into the Measurement, Distribution, and Development of Musical Capacity. Brit. J. Psychol. Monograph Supplements. Cambridge University Press, London (1948)
7. Seashore, C.E.: Manual of Instructions and Interpretations for Measures of Musical Talent. Educational Department, Columbia Graphophone Company, New York (1919)
8. Coutinho, E., & Cangelosi, A.: Musical emotions: Predicting Second-by-Second Subjective Feelings of Emotion From Low-Level Psychoacoustic Features and Physiological Measurements. Emotion. 11(4), 921-937 (2011)
9. Gabrielsson, A., & Juslin, P. N.: Emotional Expression in Music Performance: Between the Performer's Intention and the Listener's Experience.: Psychol. Mus. 24(1), 69-91. (1996)
10. Müllensiefen, D., Gingras, M., Stewart, L., & Musil, J.: The Goldsmiths Musical Sophistication Index (Gold-MSI): Technical report and documentation v0.9. Goldsmiths, University of London, London (2011)
11. Gordon, E. E.: Advance Measures of Music Audiation. Riverside Publishing Company, Chicago (1989)
12. Wallentin, M., Nielsen, A. H., Friis-Olivarius, M., Vuust, C., & Vuust, P.: The Musical Ear Test, a New Reliable Test for Measuring Musical Competence. Learn. Individ. Differ. 20(3), 188–196 (2010)

SOI MOI ...

n + n Corsino and Jacques Diennet

Ubris Studio
4, Rue Bernard du Bois
BP 62042
13201 Marseille Cedex 01 France
ubris.studio@free.fr

Created by n + n Corsino
Development: Samuel Toulouse
3D scenography: Patrick Zanoli
Sound creation: Jacques Diennet
Performance: Stefania Rossetti, Ana Teixido, Norbert Corsino
http://youtu.be/mI0MoIb5CgE



1 Introduction

See dance is to grasp the moment in several areas of representation possible. The temporal continuity of this apprehension is not measurable, but it can be deformed: it refers to a topological structure of the passage of time applied to motion perception.

The program designed by Corsino is ludic and is based on sequences that present fluid movements produced through durable artistic collaborations (the

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

dancers Ana Teixido and Stefania Rossetti, the composer Jacques Diennet, the graphic designer Patrick Zanoli) and psycho-sensory effects obtained by a long process of programming from the developer Samuel Toulouse. We no longer speak today of viewer participation, feedback, interactivity but intuitive navigation. The spectator-actor, the manipulator-player can make his scales to infinity, change wallpaper, background music, tempo, distort the image in real time ... Keystrokes are now rustling ecraniques, orders, hugs. And blow may be playing.

A sensitive navigation in harmony with the iPhone. Blow, touch, shake, push: poetic abstract kinetics of bodies and landscapes increases through the object held in the hand. Soi Moi, a work of art specifically designed as an intimate extension of oneself, offers an interactive journey through choreographic sequences and opens a new realm of creative imagination in dance.

2 Synopsis

A sensitive navigation resonates with the iPhone. Kinetics of bodies and landscapes, poetic and abstract just to increase through the tool back and specifications of the subject in hand are thereby developed through the interaction of motors.

Soi Moi, a mobile facility that gives a different perception of his own body. The iPhone becomes more user-friendly and reveals a physical sensation never felt before. Fifteen interactive sequences lasting one to two mns long form the basis screenplay. Soi Moi in the choreographed motion capture sequences plays with invisibility: the subtraction of object or partner offers unexpected physical situations. It strengthens the technical processes about when operating in the disappearance. Or more precisely in the kidnapping. Kidnapping or abduction understood as relief. Beyond the two words in the title, the construction of internal and external pressure causes some exhaust to form a tensegrity, tensile integrity, closer to the fields of biology and architecture than shamanism.

We like to think that the choreography, music and sound, set design, light and image are parallel scenarios with respect to a central theme. None is a priori worked in illustration of the other or treated as direct application. It is the same interactive mode. Mapping of the representation does not overlap with perceptual mapping of the user: they correspond in an appropriate language and a relational game resulting in a narrative form.

Related links

Last access on 20th April 2012:
http://www.liberation.fr/culture/01012390484-n-n-corsino-croisiere-virtuelle
http://www.parisetudiant.com/etudiant/sortie/n-n-corsino.html
http://www.festivaldedanse-cannes.com/IMG/pdf/com_presse_MUES_Miramar_Cannes_nov_2011_2_.pdf
http://www.yesicannes.com/yesicannes/mues_n+n_corsino.html

Author Index

Abeßer, Jakob, 567
Adhitya, Sara, 94
Ahonen, Teppo E., 474
Aramaki, Mitsuko, 257, 278

Baldan, Stefano, 437
Baratè, Adriano, 437
Barbancho, Isabel, 552
Barthet, Mathieu, 220, 492
Bernardes, Gilberto, 265
Beveridge, Scott, 508
Bi, Minghui, 102
Bogdanov, Dmitry, 618
Bravo, Fernando, 600
Brehm Nielsen, Jens, 526

Caetano, Marcelo, 287
Cano, Estefanía, 421
Chandra, Arjun, 674
Chen, Xiaoou, 70
Chew, Elaine, 634
Chordia, Parag, 344
Cochrane, Tom, 20
Coghlan, Niall, 29
Conan, Simon, 257
Correa, Debora C., 152, 466
Corsino, n + n, 695

da F. Costa, Luciano, 152, 466
De Bie, Tijl, 53
Dias, Rui, 482
Diennet, Jacques, 695
Ding, Yelei, 102
Dittmar, Christian, 421
Dixon, Simon, 395
Dolhansky, Brian, 534
Dong, Shi, 241
Dressler, Karin, 319
Dupke, André, 206

Elnusairi, Budr, 311

Farrar, Natasha, 3
Fazekas, György, 492
Ferguson, Sam, 3, 136

Gómez, Emilia, 583
Garavaglia, Javier Alejandro, 112
Giraud, Mathieu, 661
Glette, Kyrre, 674
Graepel, Thore, 357
Grollmisch, Sascha, 421
Groult, Richard, 661
Guan, Di, 70
Guedes, Carlos, 265, 482

Handelman, Eliot, 371
Haro, Martín, 544
Herrera, Perfecto, 518, 618
Hirata, Keiji, 645
Hu, Ruimin, 241

Jaimovich, Javier, 29
Jezierski, Roman, 78
Jiří Musil, Jason, 311

Kim, Youngmoo E., 186, 534
Kirke, Alexis, 457
Knapp, R. Benjamin, 29
Knox, Don, 508
Kolozali, Sefki, 220
Kronland-Martinet, Richard, 257
Kuuskankare, Mika, 94, 128, 449

Larsen, Jan, 526
Laurier, Cyril, 518
Law, Lily, 691
Le Groux, Sylvain, 160
Lee, Seungjae, 429
Levé, Florence, 661
Levada, Alexandre L. M., 466
Li, Dengshi, 241
Li, Rongfeng, 102
Li, Wenxin, 102
Liebetrau, Judith, 78
Ludovico, Luca A., 437
Lyon, Richard F., 295

Müllensiefen, Daniel, 311
Madsen, Jens, 526
Maestre, Esteban, 177
Marchini, Marco, 177
Marques, Telmo, 482
McAdams, Stephen, 45
McPherson, Gary E., 3, 136
McVicar, Matt, 53
Merazka, Fatiha, 194
Miranda, Eduardo, 457
Morrell, Martin J., 233
Morton, Brandon, 186, 534

Nagano, Hidehisa, 591
Noorzad, Pardis, 379
Nymoen, Kristian, 674

O'Hanlon, Ken, 591

Papiotis, Panos, 177
Park, Nocheol, 429
Pearce, Marcus, 395
Pennycook, Bruce, 265
Perego, Tommaso, 144
Perez-Reche, F. J., 152
Pinto, Alberto, 411
Plumbley, Mark D., 591
Prockup, Matthew, 186, 534

Refsum Jenseniuz, Alexander, 674
Rehaag, Thomas, 206
Reiss, Joshua D., 233
Rohrmeier, Martin, 357
Ross, David A., 295
Rosset, Olivier, 20
Russo, Frank A., 336

Sammartino, Simone, 552
Sand Jensen, Bjørn, 526
Sandler, Mark, 220, 492
Sarasúa, Álvaro, 518
Schmidt, Erik M., 186, 534
Schneider, Sebastian, 78
Schubert, Emery, 3, 136
Scott, Jeffrey, 186, 534
Sears, David, 45
Seo, Jin S., 429
Serrà, Joan, 544
Sigler, Andie, 371
Sioros, George, 482
Song, Yading, 395
Srinivasamurthy, Ajay, 344
Stockman, Tony, 611
Stowell, Dan, 634
Sturm, Bob L., 379

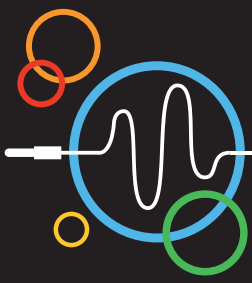
Tardòn, Lorenzo J., 552
Taylor, David, 3, 136
Thoret, Etienne, 278
Tojo, Satoshi, 645
Torresen, Jim, 674
Tran, Dieu-Ly, 45
Trochidis, Konstantinos, 45

Van Balen, Jan, 544
Velay, Jean-Luc, 278
Vempala, Naresh N., 336
Verschure, Paul F. M. J., 160
Voldsund, Arve, 674

Walters, Thomas C., 295
Wang, Heng, 241
Wang, Song, 241
Wiering, Frans, 287
Wilkie, Sonia, 611
Wilmering, Thomas, 206

Yang, Deshun, 70
Ystad, Sølvi, 257, 278

Zapata, José R., 583
Zentner, Marcel, 691
Zhang, Maosheng, 241



www.cmmr2012.eecs.qmul.ac.uk